To appear in S. Iwasaki (ed.), *Japanese/Korean Linguistics 17*. CSLI Publications. (Special section of papers from a workshop on ‘Progress in Generative Grammar’)

Should we impeach armchair linguists?

COLIN PHILLIPS
University of Maryland

1. A looming crisis?

If you believe what you read in the papers (no, not those ones - I mean journal articles, chapters, etc.), you will surely know that linguistics faces a crisis. This is because it is a field that relies on intuitive judgment data that is informal, unreliable, and possibly just plain wrong. Of course, intuitive judgments may have turned up a few facts that turn out to be reliable. But we should not take too much solace from that, because the easy observations have already been mined. A budding young linguist who enters the field today should not expect the fast facts and easy living enjoyed by his or her forbears. Instead, s/he will be forced to use increasingly sophisticated tools and methods to probe increasingly subtle facts. Armed with these tools, the New Linguist will be able to leave behind the confusions of the past and gain new insights into the nature of language.

And what is the primary tool that will replace those unreliable intuitive judgments from professional linguists?

Experimental Syntax.

That is, *lots and lots* of intuitive judgments from large numbers of people who know as little as possible about linguistics. Sometimes the judgment responses will be recorded as continuous values rather than as simple yes/no responses, allowing for fine-grained measurements and ideally a little statistical analysis. There will be graphs, and *p*-values, and effect sizes, so we'll really know What's Going On.

I am certainly prone to believe too much of what I read, but I remain to be convinced that we face a crisis (at least, not *this* crisis), and I am not sure that large-scale judgment surveys will bring the clarity that some have promised. Don't get me wrong - I love experiments, and especially experiments about linguistics. I also think that there are cases where large-scale judgment surveys can be quite informative. We have tested hundreds of speakers using this method in our own work, and we believe that it is worthwhile. However, it is important to be realistic about the nature of the problem, and the likely payoff.

It is not difficult to find complaints about the use of intuitive judgments, coming from a variety of different directions. Psychologists decry what the linguists do as irrelevant:

> Generative theories appear to rest on a weak empirical foundation, due to the reliance on informally gathered grammaticality judgments. […] A set of standards […] should be established. If these […] ideas were considered, linguistic developments might once again be relevant to the psycholinguistic enterprise. (Ferreira 2005, p. 365)

> Judgments are inherently unreliable because of their unavoidable metacognitive overtones, because grammaticality is better described as a graded quantity, and for a host of other reasons. (Edelman & Christianson 2003, p. 60)

And some linguists are inclined to concur.

> One might in fact conclude that we have not yet developed a means to evaluate empirical bases for hypotheses in generative grammar that is compelling enough to the majority of the practitioners. An evaluation of a given hypothesis thus tends to have an arbitrary aspect to it, influenced by such factors as whether or not the terms and concepts utilized are taken from a theory currently in fashion … (Hoji & Ueyama 2007, p. 2)

> Unfortunately, the findings of the experimentalists in linguistics very rarely play a role in the work of generative grammarians. Rather, theory development tends to follow its own course, tested only by the unreliable and sometimes malleable intuitions of the theorists themselves. The theories are consequently of questionable relevance to the facts of language. (Wasow & Arnold 2005, p. 495)

> Studies of usage as well as intuitive judgments have shown that linguistic intuitions of grammaticality are deeply flawed, because (1) they seriously underestimate the space of grammatical possibility by ignoring the effects of multiple conflicting formal, semantic, and contextual constraints, and (2) they may reflect probability instead of grammaticality. (Bresnan 2007, p. 75)

These are grave allegations, and so they deserve to be taken seriously. But we must be clear about the nature of the charges. The claim is not just that one finds questionable examples in linguistics papers, but that lax data-collection standards have led to the growth of theories that are based upon bogus data. If these charges stick, then we face a genuine crisis. In order for there to be a crisis, however, it would need to be the case that (i) Intuitive judgments have led to generalizations that are widely accepted yet bogus. (ii) Misleading judgments form the basis of important theoretical claims or debates. (iii) Carefully controlled judgment studies would solve these problems. Although I sympathize with the complaint that one can find many cases of poor data collection in the linguistics literature, I am not sure that any of (i)-(iii) is correct. A surprising number of the critiques that I have read present no evidence of the supposed damage that informal intuitions have caused, and among those that do provide specific examples it is rare to find clear evidence of the lasting impact of questionable judgments.

Before proceeding I should pause to clarify a couple of points for readers who may be appalled at the direction in which this argument appears to be headed. I hope that I am not seen as defending sloppy linguistic argumentation or careless construction of example sentences - I am annoyed by these as much as the next guy. I also do not mean to argue that comprehensive studies of acceptability are worthless - studies of this kind are part of the staple diet in our own work. Nor would I argue against the value of exploring linguistic questions using diverse methods - it never hurts to have a versatile toolkit. What I am specifically questioning is whether informal (and occasionally careless) gathering of acceptability judgments has actually held back progress in linguistics, and whether more careful gathering of acceptability judgments will provide the key to future progress.[1] I suspect that the real challenges lie elsewhere.

2. How reliable are intuitive judgments?

It is not difficult to dig through a few linguistics papers to find a list of questionable acceptability judgments. However, it is less easy to find cases of widely accepted generalizations that are based upon suspect data. Although the typical ‘armchair linguist’ does not systematically test his generalizations using large sets of example sentences and many naïve informants, empirical claims nevertheless undergo extensive vetting before they attain the status of ‘widely accepted generalization’. If a key judgment is questionable, this is likely to be pointed out by a colleague, or by audience members in a talk, or reviewers of an abstract or journal article. If the ques-

[1] See den Dikken et al. (2007) for interesting arguments in favor of focusing on individual judgments. That paper is part of a spirited collection of papers on the topic of grammaticality and data collection in a special issue of *Theoretical Linguistics*.

tionable generalization somehow makes it past that point, then it will still be subjected to widespread scrutiny before it becomes a part of linguistic lore.

In our lab we frequently conduct controlled acceptability judgment studies. Typically these serve as the background for a series of studies of speakers' on-line interpretation processes. The studies are easy to conduct, since the most difficult part - creation of materials - is already done for the on-line studies that they accompany. We have to run the judgment studies in order to convince skeptical reviewers that we are investigating real phenomena, but the results are rarely surprising (the same cannot be said for the on-line data, alas).[2] The sentences in (1)-(2) give representative examples. The examples in (1) test Condition C, which blocks coreference between a pronoun and an NP that it c-commands, and show that the effect is robust across a range of syntactic environments (Kazanina et al., 2007).[3] The examples in (2) test parasitic gaps, a curious phenomenon where the strong unacceptability caused by a gap inside a syntactic island is remedied by the presence of an additional gap outside the island (Engdahl 1983, Culicover & Postal 2001). Although linguists often regard parasitic gaps as marginal constructions, we found (to our surprise, quite honestly) that they are judged to be just as good as sentences without a parasitic gap.

(1) Tests of Binding Condition C (Kazanina et al. 2007)

Expt 1: 40 participants, 12 pairs of items, 5 point rating scale

a. Because last semester she was taking classes full-time while Kathryn was working two jobs to pay the bills, Erica felt guilty. [mean rating = 1.4, std err = .12]

b. Because last semester while she was taking classes full-time Kathryn was working two jobs to pay the bills, Erica felt guilty. [mean rating = 4.1, std err = .13]

[2] One class of exceptions to this is studies of 'illusory grammaticality', sentences that are fleetingly judged to be acceptable when measured during or immediately after the sentence. We have seen such effects in studies on negative polarity item licensing (Xiang et al., 2008, see also Vasishth et al. 2008), subject-verb agreement (Wagers et al., 2007), and comparatives (Fults & Phillips, 2004). A second class of exceptions involves cases of ambiguity in which the ready availability of one interpretation makes it difficult for untutored participants to recognize an alternative interpretation. We have often encountered this problem when testing scope ambiguity (cf. Goro 2007).

[3] Gordon & Hendrick (1997) argue, based on a series of acceptability rating studies, that c-command has a minimal impact upon acceptability ratings for backwards anaphora, except in environments with preposed adverbial clauses like (1a-b). The very strong contrasts found in our studies indicate that c-command plays an important role. The contrasts in (1) are also not at odds with the existence of certain semantically conditioned exceptions to Condition C, e.g., *She was about to leave when Mary noticed a letter on the doormat*, or *He then did what John always did in such situations*. See Kazanina (2005) for discussion of these cases.

Expt 2: 40 participants, 12 pairs of items, 5 point rating scale

c. It seemed worrisome to him that John was gaining so much weight, but Matt didn't have the nerve to comment on it. [mean rating = 1.5, std err = .12]

d. It seemed worrisome to his family that John was gaining so much weight, but Ruth though it was just a result of aging. [mean rating = 4.2, std err = .13]

Expt 3: 60 participants, 24 sets of items, 5 point rating scale

e. He chatted amiably with some fans while the talented young quarterback signed autographs for the kids […] [mean rating = 1.7, std err = .09]

f. His managers chatted amiably with some fans while the talented young quarterback signed autographs for the kids […] [mean raing = 3.4, std err = 0.13]

(2) Parasitic gaps: 51 participants, 24 items, 5-pt scale (Phillips 2006)

a. The outspoken environmentalist worked to investigate what the local campaign to preserve the important habitats had harmed __. [Good gap, mean rating = 3.7]

b. The outspoken environmentalist worked to investigate what the local campaign to preserve __ had harmed the annual migration. [Bad gap, mean rating = 2.1]

c. The outspoken environmentalist worked to investigate what the local campaign to preserve __ had harmed __. [Both gaps, mean rating = 3.6]

There is nothing special about these examples. They are merely representative of a broader pattern: in our experience, carefully constructed tests of well-known grammatical generalizations overwhelmingly corroborate the results of 'armchair linguistics'.

There are, of course, many cases of theoretically important judgments that are disputed. But these are typically the topic of open and lively debates, which themselves frequently lead to important new insights. For example, disputes over the analysis of long-distance reflexives in languages like Chinese and Japanese have led to new understanding about the relation between discourse and syntax (Pollard & Xue 2001, Cole et al. 2006), and controversy over the constraints on *wh-in-situ* phenomena in Japanese and other languages has led to interesting explorations of how speakers' judgments are modulated by focus and prosody (e.g., Hirotani 2005, Kitagawa & Fodor 2006, Tomioka 2007). So these are not cases of questionable generalizations that have been accepted prematurely; rather, they are cases where the facts are being actively investigated, a very healthy state of affairs.

3. Are theoretical choices based upon spurious generalizations?

Next we can ask whether there are important theoretical choices that have been made based upon disputed or sloppy judgment data. Here also I think that the answer is negative. It is a useful exercise to run through a list of well-known theoretical controversies in generative grammar and to ask, for each case, what are the key empirical disputes: transformational vs. non-transformational grammars; lexicalized grammars vs. construction-based grammars; government-binding theory vs. minimalism; relational grammar vs. phrase-structure grammar. In each case it might be possible to point to bodies of phenomena that receive elegant analyses in one theory or another, and advocates of most of these approaches would gladly list the facts that their preferred theory handles better than the competition. There are certainly disputes, but these are more likely to be disputes over what it is important to account for, rather than disputes over whether specific empirical generalizations are accurate.[4]

A reasonable objection might be that it is misleading to focus on disputes among competing grammatical 'frameworks', since these are more likely to be driven by broader philosophical considerations (not to mention aesthetics and personalities). Perhaps we can find better evidence for the damaging effects of sloppy intuitive judgments by examining theories in specific sub-domains of grammar, where the links between theory and data are more closely monitored. But this does not appear to make a difference. For example, there are important theoretical disputes in the theory of anaphora, and the past 20 years has seen a dramatic broadening of the empirical base of classical binding theory (see Büring 2005 for an excellent survey). Nobody who is seriously concerned with anaphora nowadays would deny the importance of the role of argument structure, the distinction between binding and coreference, the role of syntax vs. discourse factors, etc. However, the theoretical disputes do not appear to center on the status of key empirical data points. The disputes focus on which facts form natural classes, which are theoretically decisive, what is the scope of cross-language variation, etc. Again, all very healthy.

An example of a theoretically important domain where the facts are hotly disputed involves syntactic island phenomena. Classic syntactic theories have posited a series of formal constraints to distinguish the acceptable unbounded dependency in (3) from the various unacceptable unbounded dependencies in (4) (e.g., Ross 1967; Chomsky 1973, 1986; Manzini 1992; Szabolczi 2006).

[4] A special case may be disputes over claimed universals or 'parameters', i.e., clusters of properties that consistently co-vary across languages. However, these are disputes about cross-language sampling, and not about the reliability of judgments within an individual language.

(3) Which rumors did the press say Wilson had accused Libby of spreading __?

(4) a. * Who did the pundits wonder whether Bush would pardon ___?
b. * What did the children complain about the student who stole ___?
c. * What did the news that they had won __ amaze the team?
d. * What do Americans eat sandwiches with peanut butter and ___?

For almost as long as island effects have been known there have been prominent arguments that islands do not reflect formal constraints but rather arise from constraints on information structure (Erteshik-Shir 1973, Deane 1991) or language processing (Berwick & Weinberg 1984, Kluender & Kutas 1993, Hawkins 1999, Hoffmeister et al. 2007). This certainly qualifies as an area where the status of crucial empirical data has been challenged. However, the dispute in this case has little to do with the acceptability of the relevant sentences - on this point there is relatively broad agreement - the dispute centers on the underlying cause of the unacceptability, and whether it reflects structural ill-formedness, semantic awkwardness, or simply an overburdening of language processing resources. Teasing apart these alternative explanations is important, and not at all easy, but the debate is not a consequence of careless acceptability judgments.[5]

4. What are large-scale rating studies the solution to?

Next we can next ask whether in cases of genuinely subtle or unclear acceptability judgments our problems are likely to be resolved by using large-scale surveys.

Here, again, I suspect that large-scale judgment studies are likely to be less of a panacea than we are sometimes led to believe. Acceptability contrasts that are clear when using the much maligned ‘ask a couple of friendly linguists’ method generally remain clear when testing a large number of non-expert informants. If the larger sample makes the contrast seem less clear, this is just as likely to reflect experimenter error (misleading instructions, poorly matched examples, etc.) as distortion of the facts by linguists.

On the other hand, acceptability contrasts that are subtle or disputed in a small sample will most likely remain subtle in a larger survey. Of course, a large-scale rating survey might allow the experimenter to apply statistical tests and conclude that the subtle contrast is nevertheless statistically reli-

[5] Schütze (1996, pp. 36-39) discusses an interesting case where he argues that important theoretical choices have indeed depended on the status of conflicting judgments. Aoun et al. (1987) claim that the complementizer *that* blocks long-distance construal of *why* in *Why do you think that he left?* whereas Lasnik & Saito (1984) assume that it does not. Schütze is right that this is a troubling discrepancy that bore on a major theoretical issue of the time (ECP). What is perhaps more troubling, however, is that *both* of these analyses are now largely forgotten, supplanted by theories that have little to say about such examples.

able. But what do we learn from the small-but-reliable difference in acceptability? It certainly would be inappropriate to conclude that the sentence-type that is judged to be slightly more acceptable is well-formed whereas the other sentence-type is ill-formed. It might not even be the case that the small contrast in the group data reflects a consistent small difference at the level of individual speakers or individual sentences - the small difference might mask more systematic variability across speakers or lexical items. Much additional investigation would be needed in order to understand the source of the small contrast. In other words, the sensitivity of large-scale rating studies, which allows the linguist to identify small-but-reliable acceptability contrasts, often will merely confirm that a subtle contrast is a subtle contrast, although it could help to motivate further investigations of the subtle contrast that could lead to genuine new insights.

This does not mean that there is no value in careful testing of subtle contrasts. It is good practice to do so, and lends greater weight to any (justified) claims made based upon the subtle contrast. However, the large-scale tests typically yield few surprises, and serve merely to add force to insights developed through informal preliminary testing (i.e., the traditional approach).

Fine-grained ratings are most valuable in situations where the experimenter is interested in more complex patterns of acceptability that are clearly not amenable to introspection. Sprouse (2007, Ch. 3) presents a number of examples of this type in tests of island constraints. He reasons that if island effects can be reduced to the summed effects of two sources of structural complexity, as some have suggested (i.e., extraction is complex independent of syntactic environment; island environments are complex independent of extraction), then combined manipulation of extraction and structural contexts should yield two main effects in acceptability ratings. Sprouse consistently finds interactions of the two factors in addition to the two main effects, and argues that this provides evidence that the island constraints are not epiphenomenal. One may dispute the conclusion drawn from Sprouse's arguments - and the arguments do indeed depend on assumptions about rating scales that remain poorly understood - but it is hard to imagine how one could construct such arguments without fine-grained ratings. Similarly, arguments presented by Keller (2000) and Featherston (2005) involve complex patterns of acceptability that are probably too subtle to probe using traditional methods. This is where rating studies will prove most useful.

5. Notes on acceptability tests

There has been much recent interest in the use of new methods of quantifying acceptability. They go beyond the traditional methods of armchair linguistics, but they are fairly straightforward. Some experimental studies simply gather large numbers of binary good-bad judgments, and others ask participants to give ratings on a Likert scale, typically with 5-point or 7-point scales. But the most attention has been given to the technique of Magnitude Estimation (ME), a method that asks participants to rate the relative acceptability of a target sentence and a base sentence ('modulus' sentence), using an arbitrary ratio scale (i.e., *How many times more acceptable is this sentence than the modulus?*). ME has been argued to offer insights that other methods lack (Bard et al. 1996, Cowart 1997, Featherston 2005).

I would contend, however, that focus on the choice of different dependent measures is a bit of a distraction, and that the real action in constructing successful tests of acceptability lies elsewhere. If sufficient care is shown in the construction of suitable experimental materials, then acceptability contrasts are likely to be stable across a variety of dependent measures. (See Sprouse 2008 and Weskott & Fanselow 2008 for more thorough investigation of this issue.)

It is of course preferable to test multiple sentences and multiple informants. Ideally a Latin Square design should be used, such that experimental conditions are as closely lexically matched as possible, and such that all participants confront all item sets and all conditions (though not all possible combinations of items and conditions). In our experience, we typically obtain reliable contrasts using relatively small numbers of items and participants. It is generally sufficient to test 10-20 participants and 3-4 trials per condition per participant (multiply this by the number of conditions to find the total number of experimental item sets needed, e.g., 12-16 sets for a 4-condition study). One can, of course, test more items or participants, but that rarely makes a difference (the same is not true for reading time measures).

A crucial part of the experiment design, irrespective of the dependent measure being used, is the choice of non-target experimental items. Participants' judgments of the target sentences will vary greatly as a function of the other sentences in the study. In ME the choice of modulus sentence is obviously important. Ideally, it should have a degree of acceptability that lies roughly in the middle of the range covered by the target sentences. Of course, one should not overlook the danger that repeated comparison of different sentences to a single marginally-acceptable sentence might lead participants to evaluate the target sentences in terms of a specific salient feature of the modulus. In scale-rating studies the range of non-target sentences in the study (if there are any) also provides the background that participants use to calibrate the rating scale. If the non-target sentences are much easier or much harder than the target sentences, this is likely to compress ratings for the target sentences. The effect of this can be clearly seen in (1) above: Experiment 3 (1e-f) included examples of forwards anaphora, which are judged to be highly natural, and consequently the ratings for sentences involving backwards anaphora were compressed. Similarly, if one wants to test for subtle variations in acceptability in sentences that are not particularly bad, one should not load the study with disastrously bad sentences, as they will tend to blind participants to the contrast of interest.

Raw numbers are not particularly helpful unless one has one or more reference points to compare the ratings to. Does a mean rating of 3.7 on a 5-point scale count as acceptable? That's hard to assess for a score viewed in isolation, but if one can establish that uncontroversially well-formed sentences of the same length and number of clauses as the target receive a mean rating of 4.1, then it is reasonable to conclude that the 3.7 is a pretty high acceptability rating.

In on-line comprehension studies it is normal to include large numbers of filler items, with the aim of masking the target items and limiting the impact of task-specific strategies. In an acceptability rating study it may not be necessary or even desirable to include filler items, depending on whether the experimenter wants to draw participants' attention to the target sentences. There are no hard and fast rules here: choices should be made on a case-by-case basis.

Something that is undoubtedly important in acceptability rating studies is that participants have a very clear understanding of what they are being asked to do. This includes such basic issues as understanding the direction of the rating scale and understanding the difference between plausibility and well-formedness, or the difference between intuitive acceptability and the norms of prescriptive grammar. The participants are naïve subjects, so these things may not be obvious. It is worthwhile to work through instructions and examples with the participant and answer any questions that s/he might have about the task before proceeding. For these reasons I have grave reservations about web-delivered experiments that participants are free to run at home in their own time.[6] Even when testing students alone in laboratory testing rooms we find it increasingly necessary to remind them to turn off their mobile phones during the experiment (or at least to not answer them!). If the experiment is run in a dorm room, how confident can you be that you have your participant's full, undivided, and wakeful attention? I thought so.

Finally, it is valuable to look beyond mean ratings to look at the distribution of scores in order to gain a better understanding of participants' judgments. The results can be surprising. For example, in a study of the acceptability of backwards coreference in Japanese speakers (Aoshima et al.

[6] This concern does not undermine the potential value of web-delivered experiments, as allowed by packages such as WebExp (http://www.webexp.info). The location- and platform-independence of such packages is compatible with their use in carefully controlled settings.

2008, Exp 1A) we found that all 4 conditions yielded mean ratings of 2.9 to 3.5 on a 5-point scale, yet only around 10% of the individual scores was a '3', i.e., the score closest to the mean. Almost all judgments were 1 or 2 ('bad') or 4 or 5 ('good'), and differences in the means simply reflected shifts in the number of scores in the lower or higher categories.

6. So where is the crisis?

I have argued here that there is little evidence for the frequent claim that sloppy data-collection practices have harmed the development of linguistic theories. Does this mean that I think that all is well in theory-land? Far from it! I just don't think that the problems will be solved by a few rating surveys.

So what are the real barriers to progress in generative grammar (that was, after all, the theme of the workshop that led to this paper)?

First, we face a different kind of problem in the relation between theories and data. One of the reasons why informal data collection methods have had such a limited impact upon theory building is that there is diminishing awareness of the relation between well-known theoretical claims and their empirical motivation. Relatedly, at least in the area of syntax, there is surprisingly little engagement that cuts across major theoretical divisions in search of empirical tests. Many different sub-communities are engaged in lively internal debates, but with surprisingly limited regard for how all of the pieces fit together. Dialog between adherents of different approaches is alarmingly rare. This may in part be due to massive expansion over the past 20 years in the empirical base of linguistics: it is difficult to keep track of developments across many different areas and languages, and correspondingly daunting to try to synthesize the results into a coherent theoretical picture. But it may also be due in part to increasing loss of consensus on what the goals of linguistic theory are.

Second, I agree with many of the critics of traditional methods that there is a dire need for a deeper understanding of 'gradient' acceptability and the relation between acceptability judgments and linguistic behavior. Experimental syntax allows us to add some numbers to what is already well known, namely that acceptability comes in many shades of uncertainty. But this merely begs the question of why those shades of uncertainty exist.

One approach to gradience is to assume that speakers' ability to assign scalar acceptability values to sentences is a direct reflection of an underlying mechanism that similarly assigns scalar values to sentences. For example, Sorace & Keller (2005) view gradient acceptability as directly predicted by a grammar involving soft constraints, each of which can be violated at a cost. Similarly, Bresnan (2007) links the gradient acceptability of dative alternations to underlying knowledge of corpus probabilities.

An alternative (and very traditional) approach is to view gradient acceptability like IQ or credit rating scores, as a somewhat clumsy consequence of trying to express the output of many different properties on a single scale. Under this approach the fact that speakers can assign scalar ratings to a sentence does not indicate that any part of their language system assigns a corresponding scalar value to that sentence. Gradient acceptability may reflect the combined response to the grammar's ability to generate the sentence, to the violation of grammatical filters, to the possibility of recovering an intended meaning, to the ability to diagnose a specific constraint violation, to the availability of alternative formulations, etc. Future work should move beyond the documenting of gradience to search for ways of testing what is the underlying cause of gradience. This will likely be an undertaking where it is particularly useful to draw upon multiple techniques.

An instructive example in this regard is the famous case of constraints on *wanna*-contraction. It is widely reported that *wanna*-contraction is possible in the object question in (5a) but not in the subject question in (5b) (Lakoff 1970). The contrast is deservedly prominent in introductory courses, as it is a good example of a constraint that speakers come to know without explicit instruction. It is also cited as a constraint that children do not need to learn (Crain & Thornton 1998), although more recent evidence suggests that learning is required (Zukowski & Larsen 2007).[7]

(5) a. Who do you wanna dance with?
b. *Who do you wanna dance?

In their critique of informal judgments Wasow & Arnold (2005) present *wanna*-contraction as evidence that 'seemingly robust primary intuitions may not be shared by everyone' (p. 1483). They are certainly correct that informal questioning suggests some variability in intuitions, and it is also true that one encounters 'violations' of the constraint in natural speech, such as (6).

(6) That's the guy I wanna push *my* sled.
[Winter Olympics 1998, female commentator on burly bob-sledder]

However, systematic tests of large groups of adult speakers suggest that the constraint is indeed robust across speakers, although not without some 'noise' in specific tasks. Karins & Nagy (1993) show that *wanna*-contraction dramatically reduces the subject-extraction interpretations of potentially ambiguous sentences like *Which one would you {want to / wanna} help?* Zukowski & Larsen (2007) show in an elicited production

[7] *Wanna*-contraction is often also cited as evidence for the need for traces in the representation of unbounded dependencies, but this probably should not be counted among the more compelling arguments on that issue (Postal & Pullum 1982, Goodall 2006).

task that adults produce 5 times as many instances of *wanna*-contraction in object questions as in subject questions. Bley-Vroman & Kweon (2002) present a particularly interesting cross-method comparison, and show that although a small percentage of native speakers produce 'illicit' *wanna*-contraction in subject questions in elicited production tasks, the same speakers show almost exceptionless sensitivity to the constraint in an acceptability judgment task. It has often been suggested that the status of *wanna*-contraction varies across speakers due to the existence of 'liberal dialects' (Postal & Pullum, 1982), but clear evidence of the supposed dialects remains elusive, leading some to speculate that variable judgments might reflect intuitions that presume differing speech rates (Carden 1983).

I suspect that the moral to be drawn from *wanna*-contraction is not that linguists' judgments are misleading, but that language production reflects the interaction of multiple mechanisms, and thus is a valuable-but-imperfect window on the operations of the mental grammar.

Third, I think that it would help a great deal if more linguists were to take more seriously the mentalistic commitments to which they profess. Most generative linguists would assent to the notion that their theories should be responsive to learnability considerations, yet there has been surprisingly little exploration of how to relate current understanding of cross-language variation to models of language learning. Similarly, many linguists are happy to talk about grammar as a 'computational system' of the mind, but there is relatively little concern with the question of whether the proposed computations are actually carried out by humans. But that is a discussion for another place.

In sum, I agree with many of the critics cited above that some fundamental questions must be addressed (or readdressed) if generative linguistics is to again seize the initiative in the study of language. The perception on the outside that mainstream linguistics is becoming irrelevant is unfortunately very real indeed. However, I do not think that we should be fooled into thinking that informal judgment gathering is the root of the problem or that more formalized judgment collection will solve the problems.

References

Aoshima, S., M. Yoshida, & C. Phillips. 2008. Incremental processing of coreference and binding in Japanese. In press, *Syntax*.

Aoun, J., N. Hornstein, D. Lightfoot, & A. Weinberg. 1987. Two types of locality. *Linguistic Inquiry* 18: 537-577.

Bard, E., D. Robertson, & A. Sorace. 1996. Magnitude Estimation of linguistic acceptability. *Language* 72: 32-68.

Berwick, R. & A. Weinberg. 1984. *The Grammatical Basis of Linguistic Performance*. Cambridge, MA: MIT Press.

Bley-Vroman, R. & S. Kweon. 2002. Acquisition of the constraints on *wanna*-contraction by advanced second language learners: Universal Grammar and imperfect knowledge. Ms. U of Hawaii.

Bresnan, J. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In: S. Featherston & W. Sternefeld, eds., *Roots: Linguistics in Search of its Evidential Base,* pp. 75-96. Berlin: Mouton de Gruyter.

Büring, D. 2005. *Binding Theory*. Cambridge University Press.

Carden, G. 1983. The debate about "wanna". Evidence from other contraction rules. In *Papers from the Parasession on the Interplay of Phonology, Morphology, and Syntax*. Chicago: Chicago Linguistic Society.

Chomsky, N. 1973. Conditions on transformations. In S. Anderson & P. Kiparsky, eds., *A festschrift for Morris Halle*, pp. 232-86. New York: Holt, Rinehart & Winston.

Chomsky, N. 1986. *Barriers*. Cambridge, MA: MIT Press.

Cole, P., G. Hermon, & C.-T. J. Huang. 2006. Long-distance anaphors: An Asian perspective. In M. Everaert & H. van Riemsdijk, eds., *The Blackwell Companion to Syntax*, *vol. 3*, pp. 21-84. Malden, MA: Blackwell.

Cowart, W. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage Publications.

Crain, S. & R. Thornton. 1998. *Investigations in Universal Grammar*. Cambridge, MA: MIT Press.

Culicover, P. & P. Postal. 2001. *Parasitic Gaps*. Cambridge, MA: MIT Press.

den Dikken, M., J. Bernstein, C. Tortora, & R. Zanuttini. 2007. Data and grammar: means and individuals. *Theoretical Linguistics* 33: 335-352.

Edelman, S. & M. Christianson. 2003. How seriously should we take Minimalist syntax? *Trends in Cognitive Sciences* 7: 60-61.

Engdahl, E. 1983. Parasitic gaps. *Linguistics and Philosophy* 5: 5-34.

Erteschik-Shir, N. 1973. On the nature of island constraints. PhD dissertation, MIT.

Featherston, S. 2005. Magnitude Estimation and what it can do for your syntax: Some *wh*-constraints in German. *Lingua* 115: 1525-1550.

Ferreira, F. (2005). Psycholinguistics, formal grammars, and cognitive science. *The Linguistic Review* 22: 365-380.

Fults, S. & C. Phillips. 2004. The source of syntactic illusions. Poster presented at the 17th CUNY Sentence Processing Conference. College Park, MD.

Goodall, G. 2006. Contraction. In M. Everaert & H. van Riemsdijk, eds., *The Blackwell Companion to Syntax*, *vol. 1*, pp. 688-703. Malden, MA: Blackwell.

Gordon, P. & R. Hendrick. 1997. Intuitive knowledge of linguistic co-reference. *Cognition* 62: 325-370.

Goro, T. 2007. Language specific constraints on scope interpretation in first language acquisition. PhD Dissertation, U of Maryland.

Hawkins, J. 1999. Processing complexity and filler-gap dependencies across languages. *Language* 75: 224-285.

Hirotani, M. 2005. Prosody and LF interpretation: Processing Japanese *wh*-questions. PhD dissertation, U of Massachusetts/Amherst.

Hoffmeister, P., T. F. Jaeger, I. Sag, I. Arnon, N. Snider. 2007. Locality and accessibility in *wh*-questions. In S. Featherston & W. Sternefeld, eds., *Roots: Linguistics in Search of its Evidential Base,* pp. 185-206. Berlin: Mouton de Gruyter.

Hoji, H. & A. Ueyama. 2007. Licensed generalizations. Ms. USC & Kyushu University.

Karins, A. K. & N. Nagy. 1993. Developing an experimental basis for determining grammaticality. *Penn Review of Linguistics* 17: 93-100.

Kazanina, N. 2005. The acquisition and processing of backwards anaphora. PhD dissertation, U of Maryland.

Kazanina, N., E. Lau, M. Lieberman, M. Yoshida, & C. Phillips. (2007). The effect of syntactic constraints on the processing of backwards anaphora. *Journal of Memory and Language* 56: 384-409.

Keller, F. 2000. Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. PhD dissertation, U of Edinburgh.

Kitaagawa, Y. & J. D. Fodor. 2006. Prosodic influence on syntactic judgments. In G. Fanselow, C. Fery, M. Schlesewsky, & R. Vogel, eds., *Gradience in Grammar: Generative Perspectives*. Oxford University Press.

Kluender, R. & M. Kutas. 1993. Subjacency as a processing phenomenon. *Language and Cognitive Processes* 8: 573-633.

Lakoff, G. 1970. Global rules. *Language* 46: 627-639.

Lasnik, H. & M. Saito. 1984. On the nature of proper government. *Linguistic Inquiry* 15: 235-289.

Manzini, M.-R. 1992. *Locality*. Cambridge, MA: MIT Press.

Phillips, C. 2006. The real-time status of island phenomena. *Language* 82: 795-823.

Pollard, C. & P. Xue. 2001. Syntactic and non-syntactic constraints on long-distance reflexives. In P. Cole, G. Hermon & C.-T. J. Huang, eds., *Long-Distance Reflexives*. San Diego: Academic Press.

Postal, P. & G. Pullum. 1982. The contraction debate. *Linguistic Inquiry* 13: 122-138.

Ross, J. R. 1967. Constraints on Variables in Syntax. PhD dissertation, MIT.

Schütze, C. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.

Sorace, A., & F. Keller. 2005. Gradience in linguistic data. *Lingua* 115: 1497-1524.

Sprouse, J. 2007. A program for experimental syntax: finding the relationship between acceptability and grammatical knowledge. PhD dissertation, U of Maryland.

Sprouse, J. 2008. Magnitude Estimation and the (non-)linearity of acceptability judgments. Talk presented at WCCFL 27, UCLA.

Szabolcsi, A. 2006. Strong and weak islands. In M. Everaert & H. van Riemsdijk, eds., *The Blackwell Companion to Syntax*, *vol. 4*, pp. 479-532. Malden, MA: Blackwell.

Tomioka, S. 2007. Pragmatics of LF-intervention effects: Japanese and Korean *wh*-interrogatives. *Journal of Pragmatics* 39: 1570-1590.

Vasishth, S., S. Bruessow, R. Lewis & H. Drenhaus. 2008. Processing polarity: how the ungrammatical intrudes on the grammatical. In press, *Cognitive Science*.

Wagers, M., E. Lau, & C. Phillips. 2007. Agreement attraction in comprehension: representations and processes. Submitted.

Wasow, T. & J. Arnold. (2005). Intuitions in linguistic argumentation. *Lingua* 115: 1481-1496.

Weskott, T. & G. Fanselow. 2008. Scaling acceptability: Different measures, same results. Talk presented at WCCFL 27, UCLA.

Xiang, M., B. Dillon, & C. Phillips. 2008. Illusory licensing across dependency types: An ERP study. Submitted.

Zukowski, A. & J. Larsen. 2007. *Wanna*-contraction in typical children and people with Williams Syndrome. Submitted.