Ungrammaticality, Rarity, and Corpus Use[*]

Geoffrey K. Pullum
Radcliffe Institute for Advanced Study, Harvard University
and University of California, Santa Cruz

> "*My mother used to say that there are no strangers, only friends you haven't met yet. She's now in a maximum security twilight home in Australia.*"
> — Dame Edna Everage (character invented by comedian Barry Humphries)

1 Introduction

Geoffrey Sampson defends the extraordinary claim that there is no theoretically significant non-quantitative linguistic difference between a sentence of English and a string composed of the same words in the opposite order. Let me put that a different way. Order opposite the in words same the of composed string a and English of sentence a between difference linguistic non-quantitative significant theoretically no is there that claim extraordinary the defends Sampson Geoffrey.

The only important linguistic difference between the first and last sentences of the foregoing paragraph, under Sampson's view, is whatever separates well-trodden paths through syntactic space from those that are less travelled, or readily envisageable interpretation scenarios from those calling for a little more imagination.

It should not be imagined that Sampson is simply drawing attention to the fact that the predicate 'is grammatical' has the property of degree vagueness, like the predicate 'is bald'. Degree-vague predicates have clear cases of application, clear cases where they do not apply, and intermediate cases where applicability is an indeterminate matter. But Sampson's thesis is not that as we work to separate the grammatical from the ungrammatical we encounter an area of indeterminacy, a class of examples for which there is no fact of the matter concerning whether they are grammatical or not. It is that Chomsky (1957:13) was wrong to suggest that a syntactic analysis should "separate the *grammatical* sequences . . . from the *ungrammatical*" and study the properties of the former. His title is 'Grammar without grammaticality', and he means what he says.

It should also not be imagined that Sampson is simply drawing attention to the potential uses of a statistical model of English text that uses Good-Turing smoothing techniques to decrease over-fitting and improve predictions concerning hitherto unseen strings. Pereira (2000) makes an excellent case for the interest of such work, including the very interesting observation that using an aggregate bigram model trained on English newspaper text using the expectation maximization method, the probability of Chomsky's celebrated

[*] This work was partially supported by the Radcliffe Institute for Advanced Study at Harvard University. Many thanks to Christopher Culy, Gerald Gazdar, Mark Liberman, and Barbara Scholz for reading an early draft and providing useful advice and assistance.

grammatical but nonsensical sentence *Colorless green ideas sleep furiously* is not identical (as Chomsky 1957 asserted) with the probability of the ungrammatical *Furiously sleep ideas green colorless*. Though neither string occurs in the corpus of newspaper text used, the grammatical example shows a probability 200,000 times greater than the ungrammatical one. This is a fascinating result. But although Pereira is advocating a re-evaluation of what statistical techniques can accomplish, he does not claim that the distinction between the grammatical and the ungrammatical should be abandoned. Sampson does.

What evidence does Sampson adduce for his extraordinary claim? I will not summarize his previous presentations of his view (Sampson 1987, 1992, 1995, 2001 ch. 10); but here is what he presents for our consideration in his paper in this issue:

— one instance of John Mortimer writing *as I have made it clear to you* where Sampson (using his intuition — the rules about strictly basing our research on corpora don't apply to him!) thinks he would favor *as I have made clear to you*;

— the claimed fact (from undocumented survey work) that some people list the days of the week starting with Sunday and others start with Monday;

— a reference to the younger generation's use of *Whatever* as an utterance expressing resignation or apathy, citing some intuitions about the meaning;

— some brief references to occasional incorrectness in the intuitions of laypersons and linguists on certain points (positive *anymore*, the *a* / *an* choice, complementation with prefixed verbs);

— one anecdotal instance of an adult who once fronted the wrong auxiliary when uttering a polar interrogative; and

— some highly unconvincing remarks about metalinguistic uses of *the* (from which Sampson immediately backs off).

Apart from these, and some some puzzling side expeditions into topics like word meaning and logic that don't seem to support the main drift, Sampson has only one substantive observation to make. It is presented in the graph in his Figure 1.

The graph plots rarity of NP-internal subconstituent sequences in a 131-kiloword parsed corpus against proportion of NP tokens instantiating sequences that have equal or lower frequencies. In other words, it shows the relationship between rarity of a sequence of NP parts and proportion of NPs in the corpus that exhibit a sequence no higher than that. Sampson's conclusion from the graph is that "the proportion of all noun-phrase *tokens* accounted for by low-frequency types is quite high" and thus "constructions too individually rare for their existence to be reliably confirmed by observation would collectively form too large a proportion of what happens in the language for a grammatical description which ignored them to be satisfactory."

Now, how do we start out with the claim that individually rare constructions are jointly sufficient to make up a significant percentage of running text, and arrive at the conclusion that there is no distinction between the grammatical and the ungrammatical? Why, for example, would anyone think that because a category sequence making up an NP was rare it should be "ignored" in anyone's description? You may be able to see the logic here, but I confess that I cannot.

We have seen this graph before, in Sampson (1987), revised as chapter 10 of Sampson (2001). But the compelling response to it by Culy (1998) seems to me not to have been answered. Culy gave two stochastic context-free grammars, defining two distinct artificial languages over a single vocabulary, with rule probabilities set so that the strings of the two languages would exhibit the same frequency/coverage plot as Sampson's — one under which if sentences are constructed at random, the proportion of them covered by relatively rare types of expression is quite large. However, while one of Culy's two languages was maximal (that is, by stipulation it contained *every* string over the vocabulary), the other was a sharply restricted a proper subset thereof, with infinitely many strings excluded as ungrammatical.

Culy thus constructed an artificial situation in which the premises of Sampson's argument (the facts of the frequency/coverage plot) were true but the conclusion (the absence of a distinction between grammatical and ungrammatical strings) was false. This demonstrated the independence of the two issues. Whether the frequency/coverage plot looks like the one Sampson provides simply does not bear on whether there is a distinction between the grammatical and the ungrammatical. The same kind of frequency distribution can be found both in languages where there is a well-formed/ill-formed distinction and in languages where there is not.

Sampson makes no attempt to answer to that charge here. And in Sampson (2001, chapter 10) his only response was to ask, (i) "if real language data show such a distribution, what evidence can we have for positing a grammar . . . in preference to the simpler idea that there is no fixed distinction between grammatical and ungrammatical constructions?", and to assert that (ii) "The concept of grammaticality being controlled by a fixed set of rigid generative rules is redundant, if the data observed in practice are equally compatible with an 'anything goes' approach to grammar" (Sampson 2001:177).

The rhetorical question in (i) shifts the ground to epistemology: it asks about the nature of our evidence, a different issue from the metaphysical one of whether there is a property of grammaticality. The assertion in (ii) is a conditional, the prodosis being the claim that the data are compatible with "an 'anything goes' approach". But note that in the face of Culy's observation that the strongest case for syntax came from strictly syntactic points like the ordering of determinative (D) and noun (N) in noun phrases, Sampson declines to press an initial faltering suggestion involving discussions containing literal glossing of foreign languages (*Norwegians say 'table the'*), and admits that highly syntacticized features like article-noun order and morphosyntactic concord do determine a distinction between the grammatical and the ungrammatical in English. Conveniently for critics, then, his paper contains a rejection of its own thesis.

Why discuss matters further? For three reasons. First, Sampson is right about one thing (and I'll put it in my own way): the generative syntax community really does have a habit of using transitory personal intuitions of well-formedness as if they were invariably veridical — as if speakers had direct and infallible non-perceptual access to truths about grammatical status. I think this methodology is just as deserving of condemnation as Sampson says it is, and I will discuss it, and propose a more sensible epistemology of grammar, in section 2 of this paper. Second, I think the question of just how many strings or structures are grammatical is rather interesting, and I will present some not relevant conjectures in section 3. Third, Sampson's topic is connected to the interesting issue of the alleged problem of the "poverty of the stimulus" (how children can ever come to know facts about grammar for which the evidence from everyday usage is rare), and on one point he mentions, attestation of auxiliary-initial clauses with complex subjects. he is wrong on the facts in a way that I think deserves some airing. That will be the topic of section 4.

2 The epistemology and methodology of syntax

Looking back at the syntax published a couple of decades ago makes it rather clear that much of it is going to have to be redone from the ground up just to reach minimal levels of empirical accuracy. Faced with data flaws of these proportions, biology journals issue retractions, and researchers are disciplined or dismissed.

Take Higginbotham (1984), for example. This paper presented an argument that English is not context-free (which would have been a major result). Its evidence relied on one absolutely crucial claim: that every *such that* noun phrase modifier clause contains a pronoun anaphoric to the head noun. Barbara Partee had pointed out to the author before publication that this did not seem to be true, but the paper was published anyway. Pullum (1985) responded in print, noting that Partee's challenge was fully justified, and providing empirical backing including an attested example found by Jespersen. But Higginbotham (1985) insisted, in a confusing response, that this might not matter. It does matter (see Pelletier 1988).

The empirical facts can now be checked by reference to freely available corpora. I searched the ACL's corpus of 1987–1989 *Wall Street Journal* articles, and discovered an remarkable thing: there are only three cases of *such that* clauses functioning as modifiers in NP structure (most instances of *such that* are predicative, occurring as complement to the copula), but *none* of the three NP-modifying cases contains an anaphoric pronoun referring back to the head. Here are the NPs, with filename and line so that the context can be checked:[1]

> ... *"giant-fruited ROOF-HIGH CLIMBING TOMATOES" that get "tall as a house!" such that "A Single Slice Covers a Slice of Bread."* (w7_007:12090)
>
> ... *a driving force such that an equivalent of 81 million ounces, 29 times annual supplies, were traded on the New York Mercantile Exchange last year*... (w7_011:15135)
>
> ... *new plans on top of those guaranteed benefits, such that PBGC's guarantee is a subsidy*... (w7_096:15583)

The result, then, is devastating for Higginbotham's case: the property that must be present 100% of the time to support his argument is, according to this corpus, present zero percent of the time. Time for a retraction by the journal editor? Don't hold your breath.

Take another case, also from about twenty years ago: the judgments reported by Goodall (1987) concerning coordinations in which the coordinates are of different categories. Goodall announces on page 34 that *The bouncer was muscular and a guitarist* is ungrammatical (he gives it the '*' prefix); he accepts on page 44 the grammaticality of *Pat is either stupid or a liar* (it gets no prefix); then on page 45 he accepts *John is both crazy and a genius*, but questions *John is both crazy and a Republican* (giving it the '?' prefix), and rejects *John is both crazy and an attorney* (giving it '*'). These judgments are wildly inconsistent, as if grammaticality and acceptability of expressions with coordinate structures of the form 'AdjP *and* NP' varied unpredictably with temperature or time of day. We have no way to tell what should be predicted about further, hitherto unconsidered examples. In Monty Python's "Piranha Brothers" sketch we encounter the line: *Their father Arthur Piranha, a scrap metal dealer and TV quizmaster, was well known to the police, and a devout Catholic*. Should we call that grammatical, or ungrammatical? Goodall's how-does-it-sound-to-you-today methodology yields no basis for even a guess. His claims turn on pure, unsupported linguists' intuitions — some of his own and some from Sag et al. (1985) and one or two other sources — and there is nothing to check them against when intuitions seem to waver.

The truth about the sentences just cited, I am quite sure, is that they are all grammatical. Take Goodall's allegedly ungrammatical case *muscular and a guitarist*, for example. Google didn't actually find me a citation of that word sequence itself (the usual 'sparse data' problem), but it immediately provided *he's muscular and a pretty good mat wrestler*. And there are hundreds of other hits for similar phrases out there on the web: *lean, muscular and a first-level instructor in muay Thai*; *very muscular and a former*

[1] The examples are cumbersome and distracting in their idiosyncrasies; they show rather clearly why it should not be a goal for grammarians to illustrate every point with raw, unedited corpus examples. I have trimmed these examples down to the relevant phrases.

professional sprinter; and so on. The differences between the acquired skills of mat-wrestling, Thai boxing, professional sprinting, and guitar-picking certainly do not correlate with any syntactic contrast. That is, we surely don't want to draw a syntactic distinction between *muscular and a pretty good mat wrestler* and *muscular and a pretty good lead guitarist*. Goodall's judgments of ungrammaticality are just wrong, and could only be taken seriously because they were published in pre-Google times, and there is a tradition among syntacticians of passively accepting judgments of others rather than starting unproductive 'my dialect /your dialect' disputes.

So don't get me wrong about corpus use: I'm a convert. I think corpus-based confirmations of syntactic claims can be enormously convincing (Dalrymple and Kehler 1995, for example, is a beautiful illustration of clinching an argument with corpus data). And I don't think the how-does-it-sound-to-you-today method can continue to be regarded as a respectable data-gathering technique. Psychology gave up such methodology about a hundred years ago. For one thing, lends itself so readily to abuse. In syntax, if you want some sequence of words to be grammatical (because it would back up your hypothesis), the temptation is to just cite it as good, and probably you won't be challenged. If you are challenged, just say it's good for you, but other dialects may differ. If it doesn't sound so good, decorate the context a bit to enhance its plausibility and cite it as good anyway. Or if you need the same word sequence to be ungrammatical, fiddle with the context or the meter or some irrelevant lexical choices to make it sound a bit worse, and put an asterisk in front of it.

It is just not scientifically acceptable to go on doing syntax in this sort of way, on the basis of purported facts that are neither intersubjectively checkable nor potentially falsifiable. For quite a long time now, this head-tilting grammatical investigation, this divination by consulting the inner ear, has been discrediting theoretical syntax. The trouble is, a switch to naive or absolutist reliance on other techniques produces little improvement.

Occasionally people suggest that survey work — administering questionnaires about grammaticality — would be an appropriate empirical basis for syntax, and I do know of cases where well-conducted surveys have produced very useful results. But we should not forget that Hill (1960) had survey respondents who judged *I never heard a green horse smoke a dozen oranges* was ungrammatical until it was pointed out to them that the claim was strictly true, whereupon they switched their vote. His conclusion was that this showed something was wrong with the notion of grammaticality as a property of sentences, but I think it just shows that you can get meaningless junk out of asking people questions, and collating large quantities of meaningless survey junk is not a path to truth.

Corpus use is in a different league. I think computer searching of corpora is the most useful tool that has been provided to the grammarian since the invention of writing. Time and again I have found it to be extraordinarily important as an investigative aid. So I stand with Sampson, an inveterate corpus linguistics advocate, as a champion of this new technology and the modes of work it makes possible. But 'Grammar without grammaticality' will give corpus linguistics a bad name. That would be a real disservice to the field, and I want to ward off some of the harm it will do to have a corpus linguist spout the kind of stuff we find in this paper.

We need a more sensible set of ideas about the epistemology of grammar than Sampson's dictum that everything is possible so nothing is ungrammatical. A hint can be found in Huddleston and Pullum (2002:11):

> The evidence we use comes from several sources: our own intuitions as native speakers of the language; the reactions of other native speakers we consult when we are in doubt; data from computer corpora . . . and data presented in dictionaries and other scholarly work on grammar. We alternate between the different sources and cross-check them against each other, since intu-

itions can be misleading and texts can contain errors. Issues of interpretation often arise.

What is being suggested here is that the epistemology of grammar involves something rather like what philosopher Nelson Goodman called the method of reflective equilibrium. The useful article by Daniels (2003) may be consulted for a general account of the method, which is familiar to philosophers from applications in subfields like logic, ethics, and political philosophy. Daniels describes it in terms of "working back and forth among our considered judgments. . . , the principles or rules that we believe govern them, and the theoretical considerations that we believe bear on accepting these considered judgments, principles, or rules, revising any of these elements wherever necessary in order to achieve an acceptable coherence among them." The revisions may involve not just modification but also adding new beliefs. An acceptable coherence is one under which some beliefs "provide support or provide a best explanation for others."

Linguistics adds to this picture the actual facts of usage — the contents of corpora, for example. These lend an empirical aspect to the project. But the process of settling on a certain set of claims about the rules of grammar is still one of iterated cross-checking. The linguist tentatively formulates some proposals about the rules and check their consequences against intuition (the linguist's own, or those of other speakers, or other grammarians). Those are cross-checked against the evidence of what corpora contain. The latter may quite often change the linguist's mind about the former. Sometimes it will be clear that the current proposal about the rules should be revised. Other times a cluster of rules may start to look stable enough that it seems preferable to rethink the issue of grammaticality for certain sequences. That might mean another corpus check to re-stimulate intuitions or provide a corrective. And sometimes things may be rather more subtle: the linguist finds that a distinction was missed, and while the rule is right about one class of cases, it is wrong about the others, for which a different generalization holds.

The goal is an optimal fit between a general linguistic theory (which is never complete), the proposed rules or constraints (which are quite as conformant with the general theory as we would like), the best grammaticality judgments obtainable (which are not guaranteed to be veridical), and facts from corpora (which may always contain errors). All are revisable, at every point. The issues that arise cannot be settled by simplistic panaceas. That includes solipsistic insistence on one's own idiolect, tunnel-visioned reliance on the corpus, or dogmatic dismissal of the very distinction between what is grammatical and what is not. There are no one-answer-fits-all solutions to the problems of knowledge acquisition in our subject. As H. L. Mencken's much-quoted saying put it, "For every complex question there is a simple answer – and it's wrong."

It may well be that Sampson believes linguistics should be reconstructed on an entirely judgment-independent basis, making linguistic analysis operationalizable, even mechanizable. If so, I do not agree. I take linguistics to have an inherently normative subject matter. The task of the syntactician is exact codification of a set of norms implicit in linguistic practice. Speakers' judgments of grammaticality are, under suitably optimal conditions, an important source of evidence concerning the content of those norms, though in practice they are by no means infallible, or even largely reliable. Philosophers' discussions of how anything normative can emerge in the natural world (e.g., in ethics or aesthetics) merit attention from linguists too. A grammar defines, over an indefinitely large range, a distinction between expressions that are grammatical (even if nobody ever utters them) and expressions that are ungrammatical (even if sometimes they are uttered). Exactly how it can do this in a way that is compatible with scientific naturalism is a deep and difficult question. Much of the work of the philosopher Ruth Garrett Millikan (1984, 1993) has been addressed to this sort of question. It is not to be resolved by just asserting that nobody ever follows any rules, or insisting that either personal intuitions or frequencies of attested word sequences constitute the only linguistic facts.

3 A quantitative claim concerning ungrammaticality

On the matter of the quantitative details of just how much is or is not grammatical in natural languages, I have a specific numerical claim, as different from Sampson's as it could possibly be: I claim that *almost all* strings, whether of words or lexical categories, are *un*grammatical.

This might seem an unintelligible claim at first, since the standard view takes the set of all grammatical English utterances to be countably infinite (and oddly enough this standard view is cited approvingly by Sampson in the opening words of his paper). The set of all finite word sequences whatsoever over the English vocabulary is also countably infinite, so like any two countably infinite sets, the two have the same cardinality. But my claim can be rendered perfectly coherent in terms of finite mathematics and probability theory. What I am claiming is that for each $n \geq 1$, given any choice of a list of n word tokens (repetitions allowed), almost all of the possible sequencings of those n tokens will be ungrammatical. That claim is logically independent of whether there is a largest n or not; and it is empirical in principle, given either a sufficiently vast corpus or (contra Sampson) access to informants who can reliably (or even just approximately) report on grammaticality.

I have done nothing that could be called serious statistical research on this issue, but a little work on generating random sequences of words and trying to find grammatical possibilities for putting them together convinces me that the probability of getting a grammatical sentence using n randomly selected word tokens goes down as n goes up. Random sets of three or four words often permit a few sentences to be constructed (so many common words are usable as either nouns or verbs), but larger random sets success goes down. For example, taking every thousandth word in my recent incoming email and picking the first ten gave me the set {*among*, *and*, *level*, *made*, *of*, *terrible*, *than*, *the*, *up*, *you*}. There are 10! = 3,628,800 different possible orders in which those can be arranged, and as far as I can see none of them could conceivably be passed off as a grammatical sentence.

Of course, when we pick a random sentence and try to reorder its words, we can generally make a few other sentences. But it is usually only a few. The bulk of the $n! - 1$ other orders of the n words have to be added to the number of ungrammatical sequences. And $n!$ is a very fast-growing function (faster than x^n for any x). It seems to me that the probability of a grammatical result from a random sequencing of a random multiset of n words has a limit of zero as n goes to infinity.

Sampson's thesis is in effect simply an unsupported stipulation that the probability of getting grammatical sequences from n words is 1 for all choices of n (except insofar as he weakens his claim to some degree by admitting that it cannot be maintained where matters like agreement and article placement are concerned). It seems likely that his claim is not just wrong, but essentially the opposite end from the truth.

This is all the clearer if we consider trees (or other such graphs) rather than just strings. Suppose we construct trees at random, and label their nodes at random from some suitable vocabulary of category labels, and ask what the probability is that we will hit on something syntactically permissible. Most nodes labeled Article will not even be in an NP constituent, let alone be its left branch. Even the ones that do happen to be within an NP will have exponentially rising chances of being buried in the middle or at the end somewhere as the number of nodes in the NP rises, so the probability of a lucky hit that gets an Article in the right place falls exponentially as tree size increases.

I think the dispute here illustrates rather clearly the very simple point that if we want to make sense in linguistics *we must formalize our claims*. Something has gone wrong with the discourse when two linguists defend putatively factual claims that are diametrical opposites and the dispute continues despite a demonstration (Culy's) that the evidence presented is irrelevant to the claim defended. Sampson and I broadly agree on many topics (the shortcomings of generative grammarians' methodology; the value of corpus evidence; the potential importance of computational linguistic research; the paucity of hard evidence for the truth of linguistic nativism; and so on). It is strange for us to find ourselves unable to agree on whether essentially all word strings are grammatical or essentially none are. This is not a happy state for the discourse in linguistics to be in. And it is not going to improve until syntacticians get a lot more conversant with ways of mathematicizing their subject matter — both the application of statistics to corpora and the use of logic and algebra to formalize claims about grammatical constraints.

4 Auxiliary-initial clauses and the poverty of the stimulus

I would now like to comment on one other part of Sampson's paper, and that is the section on the corpus evidence concerning the occurrence of sentences that would provide crucial disconfirmation of the (incorrect) hypothesis that to form an auxiliary-initial clause in English you front the first auxiliary of the subject-initial counterpart. Here I think he underestimates what we can do with corpora.

Let me use 'Chomsky sentence' as a nonce term for an auxiliary-initial clause with the property that the corresponding auxiliary in the counterpart subject-initial clause is not the first (leftmost) auxiliary. For example, *Could the people who are leaving early sit near the door?* is a Chomsky sentence, because the initial auxiliary (*could*) is not the leftmost auxiliary in the declarative counterpart *The people who are leaving early could sit near the door*: *are* is the first auxiliary, and *could* is the second. (I name such sentences after Chomsky because he first drew attention to them, and has been largely responsible for spreading the belief that these sentences are too rare in language use to affect language acquisition at all, the most widely discussed 'poverty of the stimulus' argument; see Pullum and Scholz 2002 for extended discussion.)

Sampson (2002:85) expresses the view that Chomsky sentences "do not occur in spontaneous speech" as far as he has been able to ascertain. He believes it might even be the case that speakers cannot construct such sentences on the fly. He observed one instance of an adult who uttered **Am what I doing is worthwhile?* instead of the apparently intended grammatical *Is what I am doing worthwhile?* (he repeats this anecdotal observation in the current paper). There are two remarks to make about this.

First, Sampson is right that errors of this sort he anecdotally attested are very interesting; and I think it is important that Ambridge, Rowland and Pine (2005) report that in their efforts to replicate earlier work by Crain and Nakayama (1987) they found (contra Crain and Nakayama) that young children really do make errors of the sort at issue here. Sampson's observation establishes that it is not impossible for an adult to do likewise as a planning error in speech, which intriguing. All of this is compatible with the view that we do learn our native language from the evidence of the utterances we hear, and we are not guided by the automated precision of an inbuilt universal grammar module of our brains that shields us from error.

However, it would not be correct to assume that native speakers never really master the ability to express error-free interrogatives of the relevant sort in spontaneous speech. At least two Chomsky sentences can be obtained immediately by searching the *Wall Street Journal* corpus for the sequence 'Is what'. And it has been entirely overlooked in the literature that the first of those, cited in Pullum (1996), was from speech. The sentence says: *Is what I am doing in the shareholders' best interest?* It can be viewed in context at lines 2990–2992 in file `w7_003`. But it was not composed by a journalist at a keyboard; it was a spontaneously spoken example transcribed by a reporter. The full context was this:

> *Mr Tsongas says he is puzzled by such observations. "Is what I'm doing in the shareholders' best interest? Then what's the problem?"*

That settles the issue of whether Chomsky sentences occur in speech: they do. And as it happens, I have captured a few more, viva voce. As I mentioned on Language Log (`http://www.languagelog.org/`; see the archive for July 27, 2003), at about 6:24 a.m. Pacific time on 26 July 2005, the BBC World Service (relayed by KAZU, Pacific Grove, CA) played a taped segment in which an unscripted interviewee who had trained with an Islamic extremist group said that the group teaches that you have to ask yourself every day,

> *"Is what you're doing enough, or not?"*

(the ungrammatical form would have been **Are what you doing is enough, or not?*). The interviewee was certainly not reading, and it sounded as if he was speaking extempore.

I heard another such example (beginning *Is what's happening . . .*), also on the BBC World Service, on Christmas Day 2005. It was a question by a reporter that sounded as if it was made up on the spur of the moment as the interview progressed. It was clearly not read from a script.

So that is three auxiliary-initial Chomsky sentences that I have collected from speech so far through an unsystematic corpus browse and some casual radio listening. And there is something to be learned from them. It is interesting that all three (like the incorrectly formed example that Sampson noted) have subject NPs involving what Huddleston and Pullum (2002) call fused relative constructions, headed by *what*. Sentences with this property might be quite important to the acquisition issue. It would be quite interesting to know how early children hear and understand fused relatives.

It is unknown how often Chomsky sentences occur, and it is unknown how often would be enough for it to make a difference. We do not have enough evidence to decide whether children could possibly learn their native language by applying statistical inductive learning methods to what they hear, or whether they are born with innate grammatical knowledge. Answering that question will need a great deal more research (see Scholz and Pullum 2006 for a review of the relevant issues). The issue of whether most learners are likely to encounter Chomsky sentences in spontaneous speech is still open, and so is the issue of whether that tells us anything. We should not close off such open questions prematurely.

In this instance Sampson appears to underestimate the evidence for his favored view (the non-innatist one), and to underestimate what we might learn if we had larger corpora of spoken English and techniques for searching them.

5 Conclusion

Those of us who agree on the power and value of corpus methods in theoretical linguistics should continue to believe in the value of close attention to what is attested and what is not, but should take care not to follow the red herring that Sampson has trailed across our path in 'Grammar without grammaticality.' It is not sensible to abandon the distinction between the well-formed and the ill-formed, despite the well-known presence of ill-formed structures in attested material, the epistemological difficulties of syntactic investigation, and the effects of the slow process of emergent syntactic change. Sampson's paper gives us not a shadow of a reason for doubting the standard view, millennia old: that the whole point of a grammar is to tell you what is well formed in a given language, and by implication what is not.

References

Ambridge, Ben, Caroline F. Rowland and Julian Pine (2005) Structure-dependence: An innate constraint? Manuscript, University of Liverpool. Presented as a poster at the 30th Boston University Conference on Language Development, Boston MA.

Crain, Stephen and Mineharu Nakayama (1987) Structure dependence in grammar formation. *Language* **63**: 522–543.

Culy, Christopher (1998) Statistical distribution and the grammatical/ungrammatical distinction. *Grammars* **1**: 1–13.

Daniels, Norman (2003) Reflective equilibrium. Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/archives/sum2003/entries/reflective-equilibrium/`

Dalrymple, Mary and Andrew Kehler (1995): On the constraints imposed by *respectively*. *Linguistic Inquiry* 26, 531–536.

Goodall, Grant (1987) *Parallel Structures in Syntax*. Cambridge: Cambridge University Press.

Higginbotham, James (1984) English is not a context-free language. *Linguistic Inquiry* **15** 119–126.

Higginbotham (1985) Reply to Pullum. *Linguistic Inquiry* **16**, 298–304.

Hill, Archibald A. (1960) Grammaticality. *Word* **17**: 1–10.

Millikan, Ruth Garrett (1984) *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Millikan, Ruth Garrett (1993) *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.

Pelletier, Francis Jeffrey (1988) Vacuous relatives and the (non-)context-freeness of English. *Linguistics and Philosophy* **11**, 255–260.

Pereira, Fernando (2000) Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society*, **358** (1769), 1239–1253.

Pullum, Geoffrey K. (1985) *Such that* clauses and the context-freeness of English. *Linguistic Inquiry* **16**, 291–298.

Pullum, Geoffrey K. (1996) Learnability, hyperlearning, and the poverty of the stimulus. Jan Johnson, Matthew L. Juge, and Jeri L. Moxley (eds.), *Proceedings of the 22nd Annual Meeting: General Session and Parasession on the Role of Learnability in Grammatical Theory*, 498–513. Berkeley Linguistics Society, Berkeley, California.

Pullum, Geoffrey K. and Barbara C. Scholz (2002) Empirical assessment of stimulus poverty arguments. *The Linguistic Review* **19**: 9–50.

Sag, Ivan A.; Gerald Gazdar; Thomas Wasow; and Stephen Weisler (1985) Coordination and how to distinguish categories. *Natural Language & Linguistic Theory* **3**: 117–171.

Sampson, Geoffrey R. (1987) Evidence against the 'Grammatical'/'Ungrammatical' distinction. W. Meijs (ed.), *Corpus Linguistics and Beyond*, 219–226. Amsterdam: Editions Rodopi B.V.

Sampson, Geoffrey R. (1992) Probabilistic parsing. J. Svartvik (ed.), *Directions in Corpus Linguistics*, 425–447. New York: Mouton de Gruyter.

Sampson, Geoffrey R. (1995) *English for the Computer*. Oxford: Clarendon Press.

Sampson, Geoffrey R. (2001) *Empirical Linguistics*. London: Continuum.

Sampson, Geoffrey R. (2002) Exploring the richness of the stimulus. *The Linguistic Review* **19**, 73–104.

Sampson, Geoffrey R. (2006) Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory*, this issue.