On Two Recent Attempts to Show that English Is Not a CFL[1]

Geoffrey K. Pullum

University of California
Santa Cruz, CA 95064

Many of the purported demonstrations of the non-CF-ness of various languages that have appeared over the past twenty-five years (reviewed in Pullum and Gazdar 1982) have been replete with both mathematical errors and empirical shortcomings.[2] In the wake of the renewed interest in the parsability of natural languages sparked by the work of Gerald Gazdar and others (see Gazdar 1981, 1982), it has become somewhat more important to linguists to attempt to determine at least whether English – the most studied language in the era of generative grammar and the primary focus for natural language processing efforts – is context-free (CF) or not. The availability of a vast fund of information about efficient parsing of CF languages (CFLs) that could in principle be put to work on parsing English is enticing, and the idea that English and other natural languages might be a small proper subset of such a constrained and mathematically well-understood proper subset of the recursively enumerable sets is highly appealing. Chomsky (1981: 233-234) has denied this, but the denial is something of a *volte face,* for in an earlier publication (1974: 48) he remarked:

> . . . the theory of phrase structure grammar (PSG) was methodologically preferable to transformation grammar (TG) . . . . PSG is both simpler, more homogeneous, and *more* restrictive, narrower in the range of permitted grammars, than TG.

Among those who believe that English does not in fact lie within the CFLs, greater effort has recently been put into searching for subsets of English that appear to show crucially non-CF properties, and the mathematical work needed to support an argument for non-CF-ness has been conducted with much greater rigor than heretofore. Two entirely new arguments on the topic have recently appeared: one by Postal and Langendoen (in this issue) and one by Higginbotham (1984). In both, the mathematical side of the argument seems sound, great care having been taken to present all the necessary steps in the proof offered. In this paper, I shall briefly examine the claims they make. My conclusion with regard to each is that they fail to make their case because closer attention reveals that their empirical claims about English are incorrect. Moreover, both arguments concern anaphora, and both fail in ways related to that concern.

1. Postal and Langendoen on Sluicing Clauses

Postal and Langendoen (henceforth P&L) argue as follows.[3] Consider the class of English sentences meeting the schema

Joe discussed some *X* but WHICH *Y* is unknown.

The capitalization indicates contrastive stress on *which*, and *Y* is anaphorically de-stressed by virtue of reference back to *X*. Such examples are called "Sluicing" sentences (the sense suggests that, in transformational terms, an additional clause *Joe discussed* has been "sluiced" away from the position following *which Y*). P&L's claim is that, if this schema is filled out with English compound nouns from the regular set

[1] I am grateful to Barbara Partee and Ray Perrault for useful discussions, and to Terry Langendoen and Paul Postal for providing me with a copy of their paper before its publication.

[2] They continue to be cited and elaborated upon nonetheless, and statements still standardly appear in the literature of linguistics to the effect that context-free parsing cannot be used for natural languages; see Rich (183: 314-315) for yet another recent example. Gazdar (1983) provides a literature review of the topic.

[3] A refinement of the context of discussion which I will ignore here is that these authors have argued that natural languages contain sentences of infinite length (e.g., infinite coordinations, see Langendoen and Postal (1984)), which means they are not recursively enumerable, and not even sets. From this perspective, the present discussion concerns those proper subparts of natural languages that contain just finite-length sentences.

0362-613X/84/030182-05$03.00

bourbon (hater + lover)*

in the positions X and Y, a sentence will be obtained if and only if an exact string match obtains between X and Y. That is, strings like (1a) are grammatical but those like (1b) are not, according to P&L.

1a. Joe discussed some bourbon-lover-hater but WHICH bourbon-lover-hater is unknown.

b. *Joe discussed some bourbon-lover hater but WHICH bourbon-lover-lover is unknown.

If this is true, and if the set of compound nouns such as *bourbon-lover-hater* (hater of those who love bourbon), *bourbon-lover-hater-hater* (hater of those who hate bourbon lovers), and so forth, is infinite (which seems reasonable to me), then, as they show, there is a simple argument from intersection with a regular set to obtain a string copying *(xx)* language, and the conclusion is that English is not CF.

The flaw in this argument is subtle, and has to do with the relation between sentence syntax and discourse structure. Note first that the "Sluicing" construction illustrated by sentences like *Joe discussed some bourbon-lover but I don't know which bourbon lover* is not exclusively intrasentential. Dialogs like the following are encountered:

2A: Joe has been blaming the fracas last night on a certain well-known bourbon-lover.

B: Oh really? I can't imagine WHICH bourbon-lover.

Notice that in such discourses, B can interpolate additional sentences before the one containing the *which*-phrase, provided the thread of the anaphoric connection is not thereby made opaque:

3A: Joe has been blaming the fracas last night on a certain well-known bourbon-lover.

B: Oh really? You surprise me. I can't guess WHICH bourbon-lover.

Now notice that the interpolated material may even be conjoined on the beginning of the clause containing the anaphoric *which*:

4A: Joe has been blaming the fracas last night on a certain well-known bourbon-lover.

B: Oh really? I'm fairly well acquainted with the people involved, but I can't guess WHICH bourbon-lover.

This possibility spells the downfall of the empirical side of P&L's argument. By careful context construction, we can get extremely close to a legitimate context of use for exactly the sort of sentences that they rule out. Consider this discourse:

5A: It looks like they're going to appoint another bourbon-hater as Chair of the Liquor Purchasing Committee.

B: Yes–even though Joe nominated some bourbon-lovers; but WHICH bourbon-hater is still unknown.

The approach is close enough for the failure of the argument to be clearly seen; it is permissible for the antecedent of the anaphorically de-stressed constituent in a construction of this type to be in a previous sentence in the discourse, and for the anaphor relation to hold across an intervening conjunct with arbitrary content. There is no reason in principle why a clause like *which bourbon-hater is unknown* should not have a conjoined clause like *Joe discussed some bourbon-lover* intervening between it and the antecedent for its anaphoric reference, and the same holds for all other choices of compound noun. The syntax of English does not demand that the immediately preceding clause contain the antecedent for the anaphoric relationship that holds here, any more than it demands that the most recent noun phrase in the current sentence should be the antecedent for a pronoun – notice that *she* can mean *Julia* in a discourse like (6):

6. I've decided to appoint Julia. Mary wanted me to choose Kathy, but I'm sticking by my decision. She'll do a great job.

Hence sentences on the pattern

Joe discussed some X but WHICH Y is unknown.

are grammatical whether X matches Y or not (though if not, then Y will not be taken as anaphorically related to X in this construction). This means that P&L have no argument for the non-CF-ness of English.

2. Higginbotham on *such that* clauses

Higginbotham (1984) argues that English can be shown to be non-CF on the basis of the *such that* relative clause construction. This type of clause, he claims, is constrained to contain a pronoun anaphorically bound to the head. Thus he regards phrases as *the woman such that she left* as grammatical but *the man such that I saw Mary* as ungrammatical. Because of this, he reasons, the intersection of the regular language

L = the woman such that (the man such that)* she (gave (this + him) to (this + him))* left is here

with English is the following language, to be referred to as A:

{the woman such that (the man such that)^n she ((gave him to him) + (gave him to this) + (gave this to him) + (gave this to this))^n left is here | $n \geq 0$, and, reading from left to right, the number of occurrences of this never exceeds by more than 1 the number of occurrences of him}

This is shown to be non-CF by a direct application of Ogden's Lemma, hence showing that English is non-CF.

Higginbotham approaches his task with much more rigor and detail than has been customary in the linguistic literature. But as with the P&L argument, the flaws lie in the attention to detailed description of English, and moreover, relate to the treatment of anaphora. In Higginbotham's argument, it is crucial that English allows as a noun phrase any string of the form

the *N* such that *z*

where "*z* is an ordinary English declarative sentence that contains an occurrence of a third-person pronoun that does not have to be taken as having its antecedent within *z*" and "*N* is any noun that agrees properly with the pronoun in number and gender". But it also crucial that *only* if there is such an unbound and properly agreeing pronoun in *z* is the string a grammatical noun phrase. The latter assumption is plainly false. Consider this example:

7. Over many years, it has become clear that Lee and Sandy were just one of those couples such that people always reported loving her but hating him.

This contains a noun phrase of the form "Det . . . N *such that z*" where *z* contains no pronoun bound to the head noun (*couples*); yet it seems fully grammatical.

In Pullum (forthcoming) I give many more examples, varying the head nouns through a considerable range. I will not repeat all of them here, but lest anyone should think that I am using solely my own judgments here as the crucial evidence in a case of disputed data, let me point out that examples of the relevant sort can be adduced from the written English of other speakers through the written history of English. For example, a colleague found the following sentence in a manuscript under anonymous review for publication:

8. Modern linguistic theory makes crucial use of grammatical categories like 'verb', 'noun', 'subject', and 'object', such that theories of universal grammar refer to languages as being SVO, VSO, SOV, etc.

And in the prose of G. O. Trevelyan (1876: 137) one may find the following very stylish example:

9. On the 20th of February the House of Commons was called upon to express its gratitude to the Governor-General; and a debate ensued, in which the speeches from the front Opposition bench were as good as could be made by statesmen, who had assumed an attitude such that they could not very well avoid being either insincere or ungracious.

There is no pronoun referring back to the head noun *attitude* here, so it does not conform to the constraint that is crucial to Higginbotham's argument.

Surprisingly, Higginbotham knows that sentences of this general sort exist, for he sites the following two noun phrases in his first footnote:

10a. every triangle such that two sides are equal

b. the number system such that 2 and 3 make 5

and addresses the issue of how they are to be distinguished from examples he counts as ill-formed, such as

11a. every book such that it rains

b. the man such that I saw Mary

He considers but dismisses the possibility that "all of these examples are grammatical NPs, although [those in (11)] are not interpretable in any natural way, perhaps owing to the irrelevance of the sentence following 'such that' to the content of the head noun". His argument is as follows.

> First, the sentence following 'such that', even in cases like [those in (10)], is in fact *never* interpreted as closed; rather, it is interpreted, where possible, as elliptical for a sentence that is not merely relevant to the content of the head noun, but further supplies a place into which binding is possible. Thus, [(10a) and (10b)] are intuitively taken as elliptical for (iii) and (iv), respectively:
>
> (iii) every triangle such that two sides *of it* are equal
>
> (iv) the number system such that 2 and 3 make 5 *therein*
>
> Their mode of interpretation, then, not only is consistent with, but further supports, the premise employed in this article.

Higginbotham is apparently committed to the view that any sentence violating his alleged constraint will be purely an elliptical version of a longer one that observes it by virtue of containing a prepositional phrase with a pronominal NP that acts as a bound variable. But with examples I have cited, it does not even seem possible to insert extra prepositional phrases to force them to have the bound variable pronouns that Higginbotham's generalization demands.

Moreover, even if it were possible to embellish these examples with prepositional phrases to carry bound variable pronouns, this would be irrelevant to the matter at hand, since the claim at issue is about sets of strings, not their interpretations. The claim that the sentences are "taken as elliptical" is completely irrelevant. They are no more elliptical than *Kennedy was assassinated,* which lacks an agentive phrase *by someone* which semantically we might argue to be "taken" to be there.

Higginbotham goes on to remark that

> . . . there is nothing semantically odd about sentences that use NPs of the sort shown in [(11)]; for instance, (v), whose subject is [(11a)], would, if grammatical, be logically equivalent to (vi):
>
> (v) every book such that it rains is on the table
>
> (vi) either every book is on the table, or it does not rain
>
> Hence the elliptical character of [(10)], and similar examples, is a fact of grammar, for which the alternative suggestion provides no explanation.

Here Higginbotham recognized that there is no principle of logic that dictates uninterpretability, even for the bizarre cases in (11). We can assume that a noun like *book* denotes a set and a *such that* clause attached to it to make an N$'$ denotes a condition that has to be satisfied by elements of the denoted set if they are to qualify as members of the denotation of the N$'$. In a bizarre case like

(11a), the condition, if true, does not restrict the denotation of the head noun at all. In a less bizarre case like *the woman such that she left,* if the pronoun *she* is bound to the head noun *woman.* only a woman who left can belong to the denotation of the phrase *woman such that she left.* In interesting cases such as (10a), the condition is vague: *two sides are equal* could be true in many ways (the triangle is isosceles; the triangle is equilateral; the triangle is scalene but two sides of a previously mentioned quadrilateral are equal; the two sides in a recently discussed hockey game are equal in their scores; and so on). Naturally, some of these are much more likely and plausible than others in typical contexts, but clearly the vacuous-condition interpretations are consistent with the more likely ones.

In the most interesting cases, those like (7), the *such that* clause suggests a constraint on the denotation set of the head noun without explicitly giving it in the syntax. In *couple such that people always report loving her but hating him,* the *such that* clause refers to a lovable female and a detestable male but does not specify grammatical roles for these individuals exterior to the clause. The pair are referred to, however, in a *such that* clause attached to the noun *couple,* which, pragmatically, provides us with an inferred him and her to allow for the interpretation of the *such that* clause *people always report loving her but hating him* as a restriction on the reference of the noun *couple.* There is no plausibility to an account that forces this pragmatic fact into the syntax by postulating an abstract prepositional phrase to contain a suitable bound variable (**couple such that people always report loving her but hating him [of it/them]*). And even if there were, this would not bear on establishing the non-CF-ness of English given that the phrase in question is not required to appear in the string. Higginbotham is exactly right about the semantically unexceptional character of the cases in which the *such that* clause fails to restrict the denotation set of the head noun other than trivially, but that is precisely what shows he is wrong about the grammatical basis of the restriction condition.

Higginbotham's argument fails, then, because, given the evidence above that *such that* clauses do not have to contain pronouns bound to their heads, we can see that the condition regarding the relative numbers of occurrences of the words *him* and *this* in set *A* does not have to be met by members of the regular set *L* in order for them to qualify for inclusion within English. English contains not only strings like (12a) but also strings like (12b).

12a. The man such that the man such that she gave this to him gave him to this left.

b. The man such that the man such that she gave this to him gave this to this left.

Both of these, being double center-embedded, are prohibitively hard to process or to contextualize, of course, but we do not operate in such matters by attempting to render naive judgments of acceptability on extreme cases. Rather, given a clear picture of what generalizations are operative in more natural cases, we apply the familiar methodology of generative grammar and extend those generalizations to the cases where unaided intuition would fail.

If the count of *him* instances relative to *this* instances does not have to be maintained, then plainly there is no proof of non-CF-ness, since a context-free grammar is readily able to keep track of the number of *the man such that* sequences relative to the number of *gave NP to NP* sequences. It is only the additional burden of keeping the *him/this* count that allows for a proof that *A* is non-CF and thus that English is.

3. Conclusion

I think there are lessons to be learned from this admittedly negative review. It is clear that linguists have not succeeded in developing strong and predictive theories of what belongs to the domain of syntax and what belongs to semantics. Rather than attempting to discern the status of each new fact as it becomes crucial to some dispute, we ought to be developing general theories of language from which the correct conclusions follow in a principled way. Notice, in the present context, that both the arguments I have reviewed relate to the topic of anaphora. In both cases, I have presented evidence against the assumption that certain anaphoric elements are syntactically constrained to be identical and clausally adjacent to their antecedents (section 1) or to be present in the string (section 2). I suspect that it is generally true that anaphoric devices that can be controlled across sentence boundaries in discourse are never subject to any intrasentential constraint on identity or overt presence. As things stand, however, this is nothing more than a hunch. One conclusion we can draw from the present discussion is that we are in need of a general and widely accepted theory of the syntax and semantics of anaphoric devices.[4]

A second conclusion I would draw is that it is time to start applying to semantically interpreted linguistic systems the kind of mathematical analysis that so far is mostly conducted with regard to stringsets. We know little about what mathematical or computation power is inherent in particular systems that do not merely generate sets of strings but pair strings with representations of meanings that are appropriate to particular situations.[5] It would be useful to have more clear results about combined syntactic and semantic systems, since no one doubts that it is the entire mapping between structure and meaning that linguists are ultimately interested in. Moreover, it has been argued fairly convincingly that some natural

[4] See Sag and Hankamer (in press) and work cited there for some important progress in this direction.

[5] See Pullum (1984: 117-118) for one minor and nor very surprising result, observed by Len Schubert: the set of sentences assigned denotations by a semantics associated with a CF-PSG can be non-CF. Partee and Marsh (1984), stimulated by Higginbotham's paper, discuss a problem that has potential relevance here. They note that the set of predicate calculus formulae with no vacuous quantification is not a CFL, and conjecture (but are not able to prove) that it is not even an indexed language.

languages cannot be described by a CF grammar in a manner that allows a suitable syntax-to-semantics mapping to be defined (see Bresnan, Kaplan, Peters, and Zaenen (1982) on Dutch).[6]

[6] Shieber (in press), which I saw in a preliminary version after completing this paper but have not seen in its final form at the time of going to press, extends the result about Dutch to make a much stronger claim about a related language, Swiss German, namely that it has a syntactic analog of the Dutch pattern and overall is not even CF. Moreover, Culy (in press) also has evidence of a natural language (Bambara, spoken in West Africa) that appears to be other than CF. This changes the background to the current dispute a lot, of course. If Shieber and Culy are right, there are some aspects of the syntax of some natural languages that call for parsing by a device with greater than CF power. What I have said above about English remains true, of course, but although there may be general facts about how anaphora works that could have enabled us to predict this, we cannot predict it simply from the proposition that universal grammar does not allow supra-CF grammars. This makes me much less confident about being able to answer the rejoinder to this article that Langendoen and Postal publish in this issue. Although their facts, which I saw just as this article went to press, are not totally convincing (because the repeated-head-noun appositive relative construction they discuss is so awkward and unnatural even at the best of times), I now do not see why in principle they might not be right. Perhaps I *have* been speaking and writing a non-CF language all these years, and simply hadn't realized it, like Molière's M. Jourdain, who didn't realize he had native competence in prose.

References

Bresnan, Joan W.; Kaplan, Ronald M.; Peters, P. Stanley; and Zaenen, Annie 1982 Cross-serial Dependencies in Dutch. *Linguistic Inquiry* 13: 613-635.

Chomsky, Noam 1974 Interview with Herman Parret. In: Parret, Herman, Ed., *Discussing Language.* Mouton, The Hague, Holland: 27-54.

Chomsky, Noam 1981 Untitled comments in reply to a question from Henry Thompson. *Philosophical Transactions of the Royal Society of London B* 295: 277-281. [Reprinted in *Psychological Mechanisms of Language.*Society and the British Academy, London, 1981. Page reference to the original publication.]

Culy, Christopher forthcoming The Complexity of the Vocabulary of Bambara. *Linguistics and Philosophy.*

Gazdar, Gerald 1981 Unbounded Dependencies and Coordinate Structure. *Linguistic Inquiry* 12: 155-184.

Gazdar, Gerald 1982 Phrase Structure Grammar. In: Jacobson, Pauline and Pullum, Geoffrey K., Eds., *The Nature of Syntactic Representation.* D. Reidel, Dordrecht.

Gazdar, Gerald 1983 NLs, CFLs and CF-PSGs. In: Sparck Jones, Karen and Wilks, Yorick, Eds., *Automatic Natural Language Parsing.* Ellis Horwood, West Sussex, England.

Higginbotham, James 1984 English Is Not a Context-Free Language. *Linguistic Inquiry* 15: 119-126.

Langendoen, D. Terence and Postal, Paul M. 1984 *The Vastness of Natural Languages.* Basil Blackwell, Oxford.

Partee, Barbara and Marsh, William 1984 How Non-Context-Free Is Variable Binding? Presented at the Third West Coast Conference on Formal Linguistics, University of California, Santa Cruz.

Postal, Paul M. and Langendoen, D. Terence 1984 English and the Class of Context-Free Languages. *Computational Linguistics* 10(3-4): 177-181.

Pullum, Geoffrey K. 1984 Syntactic and Semantic Parsaility. In: *Proceedings of Coling84.* Stanford, California: 112-122.

Pullum, Geoffrey K. forthcoming *Such That* Clauses and the Context-Freeness of English. *Linguistic Inquiry.*

Pullum, Geoffrey K. and Gazdar, Gerald 1982 Natural Languages and Context-Free Languages. *Linguistics and Philosophy* 4: 471-504.

Rich, Elaine 1983 *Artificial Intelligence.* McGraw-Hill, New York, New York.

Trevelyan, George Otto 1876 *The Life and Letters of Lord Macaulay, Vol. II.* Logmans, Green and Co., London.

Sag, Ivan A. and Hankamer, Jorge 1984 Toward a Theory of Anaphoric Processing. *Linguistics and Philosophy* 7: 325-345.

Shieber, Stuart forthcoming Evidence Against the Context-Freeness of Natural language. *Linguistics and Philosophy.*