

TOPIC . . . COMMENT

Footloose and context-free

It was not an isolated believe-it-or-not coincidence when a Cambridge mathematician (Adams) and a Paris mathematician (Leverrier) both predicted the discovery of Neptune at the same time through similar but entirely independent calculations of Uranian orbit wobble. Similar things happen all the time. Ideas often seem to be hanging from the tree of science like ripe fruits ready to fall, and several hands may grasp at the bough simultaneously.

A number of things are necessary to make it nonetheless possible to identify the discoverer(s) of a given truth: careful record-keeping about research activity, fair and efficient management of the peer-review and research publication enterprise, and above all, generally accepted standards about what constitutes a result. If things that have not by any stretch of a disordered imagination been demonstrated are claimed to have been demonstrated, clearly it will be hard to establish later that a given person discovered a given thing. If standards of evidence are set ad hoc to ensure rhetorical victories over critics, and alliances determined more by sociological groupings than by problems shared, there is little hope of being able to look back and see progress.

Let me give a case history. It is not pretty; in fact, it is a mess, but we must face our world as it really is.

By 1985 it had become clear that not all natural languages are context-free. In 1955 the question could not have been formulated, because context-free languages (CFLs), and the context-free phrase structure grammars (CF-PSGs) that by definition generate them, had not yet been defined. So the discovery that certain natural languages are not CF occurred some time between 1955 and 1985. Yet although the result in question was widely believed true by the middle sixties, it does not seem to have been validly and publicly shown to be true until quite recently, and the question of who deserves the credit is a morass of unclarity.

The question was first formulated by Noam Chomsky. In his 1956 paper 'Three models for the description of language' [*I.R.E. Transactions on Information Theory*, Volume IT-2, pp. 113-123], he reported that he didn't know the answer. During the wave of activity which was the start of the whole field of formal language theory (now an important part of theoretical computer science) a number of people started looking for the answer. Several early efforts are reviewed in R. T. Daly, *Applications of*

the Mathematical Theory of Linguistics [Mouton, The Hague, 1974] and G. K. Pullum and G. Gazdar, 'Natural languages and context-free languages [*Linguistics and Philosophy* 4 (1982), 471–504].

Paul Postal gave a non-CF-ness argument for Mohawk in his 1962 Yale University doctoral dissertation [*Some Syntactic Rules in Mohawk*; Garland, New York, 1979], and later in a 1964 paper ['Limitations of phrase structure grammars,' in J. A. Fodor and J. J. Katz, eds., *The Structure of Language*, Prentice-Hall, Englewood Cliffs; pp. 137–151], and Chomsky proposed one himself ['Formal properties of grammars,' in R. D. Luce et al., eds., *Handbook of Mathematical Psychology, Vol. II*, Wiley, New York; see pp. 378–9]. Like most of the early arguments, these had both formal and empirical failings. Although suggestions were made for patching some of them up (see especially D. T. Langendoen's paper in *Studies in Descriptive and Historical Linguistics: Festschrift for Winfred P. Lehmann*, ed. by P. J. Hopper [Benjamin, Amsterdam, 1977; pp. 159–171]), the case still seemed to Gazdar and me not to have been made.

In one argument we discussed, neither the mathematics nor the facts are in dispute, yet the issue still seems hard to resolve. Arnold Zwicky claimed in 1963 ['Some languages that are not context-free,' *Quarterly Progress report of the Research Laboratory of Electronics, MIT*, 70, 290–293] that natural language number systems are not CF. What he showed entails that if (say) *zillion* is the highest monomorphemic number-name in English, the set of all number names will be non-CF, because in a well-formed number name the zillions follow the zillion zillions, which follow the zillion zillion zillions, and so on, and a CF-PSG cannot handle this pattern if length of strings is unbounded (as it surely must be with number names). Gazdar and I took the view that this is merely a fact about the number system. We have to be taught it in math classes at school, and we do not acquire it with our language *per se*. On odd dates I still think this is right, but on even dates I think the argument has been unjustly overlooked. Zwicky thinks we were correct to dismiss it, but maybe he is wrong and it was the first valid argument that English is non-CF. The problem here is that we are not entirely sure what is a fact about a language and what is a fact about the culture associated with it (more on this below).

The most fascinating new material uncovered during the recent debate about CF-ness was the Dutch construction first discussed in this context by M. A. C. (Riny) Huybregts in 1976 ['Overlapping dependencies in Dutch,' *Utrecht Working Papers in Linguistics* 1, 24–65]. Dutch subordinate clauses with meanings like 'Al saw Bo make Cy let Di help Ed shave' come out (or can come out) with the word order *Al Bo Cy Di Ed saw make let help shave*. This rather surprising fact leads to the conclusion that this

construction in Dutch exhibits what has become known as a cross-serial dependency: the n th verb takes the n th NP as its direct object. (If you believe in small clauses, this is incorrectly put, and I have no idea what superficial structure you might assign; but then, if you believe in small clauses you probably eat steak with a spoon. Even Edwin Williams doesn't believe in small clauses.)

What Gazdar and I maintained about the Dutch case was that the purely syntactic facts seemed unproblematic. All the grammar had to do was provide the right number of NPs to go round – one per verb, in the simplest case. A context-free grammar could easily do that. However, to show this, we used grammar fragments whose chances of being associated with valid semantic rules were similar to Mr T's chances of being offered an honorary doctorate at Harvard. There was clearly something disturbing and potentially very relevant to the issue afoot, for we could not exhibit a context-free grammar which both generated Dutch and seemed likely to be able to support a semantics. (This is in essence the point that Bresnan, Kaplan, Peters and Zaenen [*'Cross-serial dependencies in Dutch,'* *Linguistic Inquiry* 13 (1982), 613–635] pursued, though they put things in terms of syntactic motivation for tree structures.) Huybregts had put his finger on the reason why Dutch was a problem, but Dutch didn't quite allow for his argument to be completed.

Now the plot quickens its pace a bit. (It should, I hear you cry; nothing has happened yet.) At a conference in Brussels in June 1983, Tom Wasow was told by Richie Kayne that Swiss German had a similar word order pattern to that of Dutch in clauses of the relevant sort, but also had, like standard Germany, certain verbs which took a visible dative case on their objects. That would mean there could be a morphologically indicated syntactic link between the n th verb and the n th NP, which would be very likely to allow a non-CF-ness proof to be constructed. As luck would have it, a native or near-native speaker was available right there at the conference: Henk van Riemsdijk has mother-tongue knowledge of Swiss German. Wasow had a lot of trouble pinning van Riemsdijk down for even a cursory interview; van Riemsdijk was very busy, and apparently uninterested in the issue at hand. Wasow did not get very far with the investigation, but by the end of the conference he had extracted one example which at least suggested there was a case to answer.

In August 1983, after having been told of this by Wasow, Stuart Shieber, a computer scientist at SRI International in Menlo Park, California, set off on a trip to Europe to attend a conference in Switzerland. Finding himself in Zurich, he started working with informants in that area to see what he could find out about the crucial syntactic property-cluster Swiss German was reputed to have.

Meanwhile, in western Mali, a hundred drought-stricken miles southwest of Timbuktu, Christopher Culy, a recent linguistics graduate who had decided to do two years in the Peace Corps before embarking upon graduate school, was commencing lessons in the language of the area, Bambara. While working on the language, he came upon something that took him back to his classes in linguistics and mathematics at Stanford. The device Bambara uses for forming an expression meaning 'whatever dog' is reduplication of the noun-stem meaning 'dog' with an *o* separating the two halves. But, Culy noticed, the construction appears to work the same for noun stems of any length, including compound nouns with internal syntactic complexity. It looked, therefore, as if the Bambara lexicon had an infinite subset with the form *xx*, where *x* could be of any length. This could be used to argue that Bambara as a whole was not CF.

On August 24, 1983, Culy wrote a letter to Tom Wasow and Ivan Sag, his former teachers at Stanford, describing the Bambara situation and sketching an argument that it made Bambara non-CF. Wasow wrote back encouraging Culy to construct the argument explicitly and write it up for publication. Culy proceeded to do this, though there were great difficulties, one being the mail delays involved in communicating with Stanford, and another being that as a government agency employee, Culy had to get permission from the Peace Corps to publish.

Whether Shieber had his results on Swiss German by the time Culy's letter was mailed from Mali is not clear to me. But a further strand complicating the story must now be mentioned. Considerably earlier, in 1981, it had occurred to Alexis Manaster-Ramer that one could develop a non-CF-ness argument using the contemptuous reduplication of Yiddish-influenced English (transformation, *schmansformation*). In the spring of 1983 he presented his argument at the Chicago Linguistics Society's Regional Meeting, and in the Summer, with Culy in Mali and Shieber in Switzerland, it was published in the CLS proceedings volume for the year ['The soft formal underbelly of theoretical syntax,' *CLS 19*, 256–262].

Now, to me, the construction Manaster-Ramer refers to looks like a game one plays *with* the language rather than a construction *within* the language (the difficulty of knowing what's language and what's culture again), and thus I was not immediately impressed with the argument. Moreover, the paper is distinctly equivocal about whether a valid non-CF-ness argument exists (see pp. 259–261). Manaster-Ramer does note that similar reduplication constructions are found in several other languages, but he does not give any examples. Manaster-Ramer's CLS paper could be seen as a prior publication of essentially Culy's point, though it isn't entirely clear.

Culy and Shieber both wrote papers and submitted them to *Linguistics and Philosophy* sometime in the first half of 1984. Culy's was submitted first, but owing to various delays, Shieber's paper was accepted first. Eventually both papers appeared in the same issue of *Linguistics and Philosophy*.

But before they did, in the spring of 1984, a wild card was played. Out of the blue, a paper by James Higginbotham appeared in *Linguistic Inquiry* [15, 225–234]; its title was, 'English is not a context-free language.' It cited none of the previous literature on the subject; one would have thought that Higginbotham had invented the issue on his own. The argument it offered involved applying a mathematical result known as Ogden's Lemma to a set of strings involving the relative clause-like construction with *such that*. The crucial empirical premise was that in an NP of the form 'Det N *such that* S', S must contain a pronoun of the right number, person, and gender to refer back to N. It had already been pointed out to Higginbotham before he published that this didn't really appear to be true (phrases like *any triangle such that two sides are equal* had been cited by Barbara Partee), but he was not deterred, and dismissed such examples in a footnote as ungrammatical but interpretable as ellipses (p. 229, n.1).

I published a response arguing that Higginbotham was entirely wrong about the facts [*Linguistic Inquiry* 16 (1985), 291–298], and he replied indignantly in the same issue of *LI* that I was completely wrong about him being wrong [298–304], and naturally I believe that he is completely wrong about me being wrong about him being wrong. (These things tend to drag on; in future work, Higginbotham will argue that my eyes are too close together, and I will argue that on the contrary, his head is too round.) But if he were right about what is and is not grammatical, then he would be the first person to have published a valid demonstration that natural languages are not context-free in a refereed journal, because it was 1985 before Shieber and Culy saw their papers in print.

Before I try to draw a moral from this historiographical chaos, let me point out what happened when a situation with some similar characteristics arose in another discipline, mathematics. In 1985, no less than five distinct groups of mathematicians hit on essentially the same result: a beautiful new way of characterizing knots in terms of polynomials. As Ivars Peterson reports in *Science News* 128.17 (October 26, 1985), p. 266, the *Bulletin of the American Mathematical Society* saw to it that the results were amalgamated, and a single paper with six authors was published, with summaries of the proofs achieved by four of the five groups (the fifth group, two Polish mathematicians, missed out because of mail delivery delays). All mathematicians agreed that the five groups had achieved

essentially the same result, and no attempt was made by any of the mathematicians involved to claim priority over the others in the discovery.

Linguists seem not to behave so well. The chaotic thirty-year history of our efforts to decide whether there are non-CF natural languages has no air of clean professionalism about it. And although Ivars Peterson wrote a report on the Culy and Shieber results in *Science News* [128.20 (November 16, 1985), 314–315], his report was promptly attacked twice in the letters column: Robin Ault of Newtonville, Massachusetts [128.24 (December 14, 1985), p. 371] bumbled incoherently that languages were really finite (and was in turn incoherently attacked by Gary R. Lavine [129.4 (January 25, 1986), p. 57]; and Michael Kac [129.2 (January 11, 1986), p. 19] protested that Manaster-Ramer had priority over Culy and Shieber.

The non-CF-ness result itself, Chomsky has repeatedly told us, is of little importance. But then hardly anything in linguistics is important, in a way: if invalid arguments or incorrectly substantiated results are reported in linguistics, society suffers no particular ill, whereas if the same thing happens in marine toxicology, we eat poisoned fish. Probably the most interesting thing about the whole debate is the view it affords of how linguists do business.

I find myself feeling chastened by the words of Columbia University mathematician Joan Birman (quoted by Peterson) about the polynomial knot results:

I felt very proud of mathematicians for the nice way that those competing announcements were handled. It had the potential for a big argument, but there was none.

For those who share my feeling that we linguists, in our disunity, may not have left quite such a good impression, a different epigraph: the closing lines of D. H. Lawrence's poem, *Snake*.

And I have something to expiate:
A pettiness.

Received 11 March 1986

Cowell College
University of California
Santa Cruz CA 95064
U.S.A.

GEOFFREY K. PULLUM

Note

The views expressed in TOPIC . . . COMMENT are those of the author. They should not be construed as representing either the editor or the publisher of *NLLT*.