On Formalization and Formal Linguistics

(From Professor Noam Chomsky)

Dear Editor,

In *NLLT* 7,1 (1989), Geoffrey Pullum laments the impending doom of formal linguistics in favor of the "gentle, vague, cuddly sort of linguistics" that he feels I have been advocating since a shift of opinion that he traces to 1979. His account, however, is based on misreading and serious misunderstandings. I will keep to the latter.

Two interrelated issues lie in the background: the status of formal linguistics and of formalization. A few comments follow on each.

Pullum quotes approvingly a passage from my *Syntactic Structures* (1957, henceforth *SS*) on the value of replacing "obscure and intuition-bound notions" by "precise and technical development of linguistic theory," claiming that in the post-1979 heresy, I "flatly reject" my 1957 position. On the contrary, I have never questioned it, though a few years later I came to realize that the project under discussion – namely, the one pursued in *Logical Structure of Linguistic Theory* (1955–6, *LSLT*) – was premature and far too ambitious.

One of the "obscure and intuition-bound notions" that should be clarified or eliminated is *set of well-formed (grammatical) expressions (E-language*, in the terminology of my *Knowledge of Language* (1986, henceforth *KOL*)). Though unproblematic (by stipulation) in the theory of formal languages, the notion remains obscure, perhaps lacking any empirical status, for natural language. In *LSLT*, a notion of E-language is indeed proposed, but as a high-level construct within the theory of *strong generation* of structural descriptions (SDs) (specifically, within the theory of derived constituent structure). Strong generation is the basic notion; *weak generation* of a designated set of strings (an E-language) is defined as an auxiliary concept, of marginal significance at best. The *LSLT* proposal was soon recognized to be faulty, and in my *Aspects of the Theory of Syntax* (1965), it is assumed that a generative grammar (an *I-language* in the terminology of *KOL*) strongly generates an SD for every expression. The only notion of E-language proposed is the trivial one: every I-language L weakly generates the fixed E-language E, where E is the set of all valid phonetic representations; each of these is assigned an SD by L. The set of SDs strongly generated (the *structure* of the language, in the suggested terminology) is a nontrivial set associated with L; the string set generated is not. See pp. 30f., and for further discussion, *KOL*, chapter 2, particularly note 17.

Natural Language and Linguistic Theory 8: 143–147, 1990.

As is further discussed in *Aspects*, even if some notion of E-language can be constructed for natural language, and if it turns out that linguistic theory (universal grammar, UG) permits (weak) generation of all recursively enumerable E-languages, the fact would be of no apparent empirical significance for learnability (60f.). There was no need to add that weak generative capacity has no apparent relevance to parsability, because nothing had yet been claimed in this regard.

Pullum defines "formal linguistics" (which he hopes to save) as the study of E-languages and the I-languages that weakly generate them. This definition raises two problems: (1) empirical significance; (2) arbitrariness.

As for (1), evidently, if Pullum expects this study to be of any empirical significance, he will have to offer some indication, however vague, of what an E-language is; I am aware of no proposal since *LSLT* (the definability of some notion in the theory of formal languages is, of course, beside the point). Note that these problems do not arise as UG is understood in *LSLT*, *SS*, and *Aspects*, or in the post-1979 work that retains the relevant assumptions.

Turning to (2), it is true that as pursued, formal linguistics largely concerned itself with weak generation of E-language, for basically two reasons: the background in metamathematics and automata theory; the fact that these are simple notions, readily investigated by available tools. But formal linguistics is not *defined* in this manner. Rather, it is the study of formal models in abstraction from application; it is formal *linguistics* rather than some other branch of mathematics insofar as one can show (or one hopes) that these models have an application in the study of natural language or provide, however indirectly, some insight into properties of natural language. Plainly, formal linguistics as such does not rest on weak generation, nor is it restricted to weak generation even if this notion is defined (nontrivially) in some formal theory. Some of the earliest results had to do with strong generative capacity (the relation between context-free grammars and pushdown storage automata, for example), and the study of E-language was suggestive insofar as it shed some light on strong generation (e.g., the study of self-embedding and copying). Current work in complexity theory (see Edward Barton, Robert Berwick, and Eric Ristad, *Computational Complexity and Natural Language*, 1987) is particularly clear in bringing out the central importance of an I-language perspective and the (at best) marginal character of weak generative capacity if complexity problems are to be seriously studied.

Pullum claims that "no sense was ever supplied to the notion of a grammar that...does not characterize any [E-language]." On the contrary, work in generative grammar from *LSLT* has proceeded in terms of strong generation, with E-language a marginal and derivative notion at best; and the formal study of grammar (I-language) can readily be developed in terms of strong generation

alone. The latter point is clear even in the work that Pullum cites as his model. Thus he cites three conditions from a study by Robert Stoll as "non-negotiable...for formal theories of grammar... ." These conditions state only that there must be a well-defined notion of I-language, structural representation (graph, diagram, etc.), and generation of such representations, with no requirement that E-language exist (except trivially, as in the sense of *Aspects*).

Pullum also believes that my post-1979 reiteration of these points leads to elimination of "the defining idealization of formal linguistics: the idea that [E-languages] are abstractly definable and can be studied in isolation from biological or biographical facts about their speakers." This is his interpretation of my statement that I-languages might "characterize [E-languages] that are not recursive or not recursively enumerable, or even...not generate [E-languages] at all without supplementation from other faculties of mind." There are multiple confusions in his interpretation of this accurate statement. First, Pullum fails to understand that this post-1979 statement merely reiterates standard assumptions of his "good old days." Second, these earlier assumptions are, so far as we know, quite accurate. Contrary to Pullum's belief, E-language has no particular significance for formal linguistics. As for more realistic models based on strong generation of structures by I-language, the project of studying them in isolation from biological or biographical facts is unrelated to the status of E-language (if any). Furthermore, the project is dubious if Pullum's remark is intended to suggest that there should be some "pure" study of language isolated from discoveries about acquisition, use, and physical mechanisms.

The issues are far from academic. It is well-known that any 2-category partition of expressions will undercut much of the most significant linguistic work. The differential effects of ECP, subjacency, selectional constraints, etc., are far more revealing than any division into well- versus ill-formed, and bear directly on central principles of UG. In contrast, the point of a [± WF]-dichotomy remains obscure, even if it can be established in some nonarbitrary fashion. Suppose that Jones has the I-language *L*, some variety of English. As far as is known, it is meaningless to ask whether a weak *wh*-island violation or such an expression as "misery loves company" is, or is not, a member of the E-language weakly generated by *L*; and nothing would follow from a discovery (or stipulation) one way or another. These expressions have their status, determined by *L*; they are parsable, appropriate in certain situations, have a definite meaning, etc. *L* also provides an interpretation for an utterance of Japanese. In fact, one could learn a good bit about *L* by determining how Jones interprets such expressions. There are further questions here, but they go beyond the issues at hand.

Turning to the role of formalization in linguistics, like Pullum, I see no need to qualify the statement he quotes from *SS*. Theories should be formulated

clearly enough, and observations firmly enough established, so that inquiry can proceed in a constructive way. Beyond that, experiments can be carried out more carefully and theories made more precise, but the burden of proof is on those who consider the exercise worth undertaking.

Sometimes the burden can be met: inquiry is advanced by better data and fuller formalization, though the natural sciences would rarely if ever take seriously Pullum's injunction that one should make "a concerted *effort*" to meet "the criteria for formal theories set out in logic books." At one time, there was some marginal interest in the project; Woodger's attempted formalization of biology is a well-known case, forgotten, because empirical consequences were lacking. Even in mathematics, the concept of formalization in our sense was not developed until a century ago, when it became important for advancing research and understanding. I know of no reason to suppose that linguistics is so much more advanced than 19th century mathematics or contemporary molecular biology that pursuit of Pullum's injunction would be helpful, but if that can be shown, fine. For the present, there is lively interchange and exciting progress without any sign, to my knowledge, of problems related to the level of formality of ongoing work.

True, work should be clear enough so that it *could* be formalized further if there is some reason to do so. The point can be illustrated with Pullum's two references to my *Barriers* (1986), one of which he finds a "total baffler," the other "worse than ever." In each case, what he cites is completely straightforward, though the monograph does presuppose some literacy in linguistics and logic.

The "total baffler" is a clause from a definition of "dominance," which reads as follows:

> α is dominated by β if it is dominated by every segment of β.

A footnote, which Pullum finds meaningless, observes accurately that further formalization is straightforward in terms of the notion *occurrence*, an elementary notion formalized in a linguistic context (borrowing a device of Quine's) in *LSLT* and an earlier article of mine in the *J. of Symbolic Logic* (1953). Pullum finds the definition circular and "casually bungled," but it is unproblematic. First, we define domination by a segment. Categories are defined in terms of segments, and we then define domination by a category as domination by every segment of this category. There is no more "circularity" than in any standard recursive definition.

The statement that Pullum finds "worse than ever" is that quite commonly, "there is no point in specifying one or another of the possible options in detail." The reference in this truism is to X-bar theory. To formalize one or another version is a straightforward exercise, but apparently of no more value than

Woodger's, because it would require decisions that are arbitrary; not enough is understood to make them on principled grounds. The serious problem is to learn more, not to formalize what is known and make unmotivated moves into the unknown. Those who think otherwise may be right, but it is their task to show it.

One should not be misled by the fact that computer applications require such moves. Throughout history, those who built bridges or designed airplanes often had to make explicit assumptions that went beyond the understanding of the basic sciences. Sometimes such steps also prove useful for advancing understanding, sometimes not. There are no doctrines on the matter.

Leaving aside a series of further misunderstandings and errors, the basic points seem simple enough. The post-1979 shift that Pullum perceives is imaginary and the lethal consequences, nonexistent. Formalization may sometimes be a valuable tool and the study of formal linguistics retains its interest, though not under Pullum's arbitrary constraints. The notion of E-language appears to be of marginal significance at best and may have no empirical interpretation at all, as assumed in the earliest work in generative grammar. Pullum's conclusions are based on serious misunderstanding throughout, and his concerns are as groundless as his fevered charges.

Cambridge, MA
August, 1989

NOAM CHOMSKY