Linguistics for Computability Theorists
Three Tutorial Lectures

Geoffrey K. Pullum

School of Philosophy, Psychology and Language Sciences
University of Edinburgh
Dugald Stewart Building, 3 Charles Street
Edinburgh EH8 9AD, Scotland (UK)
Tel.: +44-131-650-3603
gpullum@ling.ed.ac.uk

These three tutorial lectures assume a basic acquaintance with mathematics, logic, and theoretical computer science (including computability theory), but no prior knowledge of linguistics. The focus will be on syntax, because it is the least familiar to non-linguists (people generally know quite a bit more about pronunciation and meaning than they know about grammar), and because it offers interesting possibilities for applying computability theory.

1. Theoretical syntax over the past fifty years, particularly in the USA, has been almost entirely dominated by the idea that languages are sets of objects—typically strings, but more generally graphs—and grammars are finite definitions of the characteristic functions of those sets. This is known to linguists as the **generative** conception of syntax. Linguists have not adequately appreciated how directly it stems from the work of Emil Leon Post [4] on formalizing proof theory. Generative grammars are basically just special cases of Post production systems, utilized as he utilized them — to provide recursive definitions of computably enumerable (CE) sets.

However, whereas Post concentrated on systems expressive enough to define any CE set whatever, linguists have been interested in classes of generative grammars with much lower expressive power. The familiar Chomsky Hierarchy of generative grammars and languages [2] has the CE languages (type 0) at the top, with the context-sensitive (CS, or type 1) below that, and the context-free (CF, type 2) below that, and the regular or finite-state (FS, type 3) at the bottom. But linguists have been interested in a much larger range of classes than that. There are many families of stringsets properly included in the regular sets, and some interesting families lying between CF and CS.

Some attention was paid in 20th-century work to the issue of where natural languages might fall in the Chomsky Hierarchy. This question has little interest in itself, but it is a prerequisite to work on various computability questions that naturally arise in connection with human languages:
— generative capacity of various linguistic formalisms;
— decidability of properties of classes of grammars;
— decision problems about grammars and languages (emptiness, universality, equivalence, etc.);
— the recognition problem for human languages;

— the issue of language 'learnability' (in the guise of the much more specialized notion 'identifiability in the limit from positive instances' as applied to classes of stringsets).

A selection of results in these areas will be surveyed.

2. Despite the insights that have been gained from work on generative grammars and the useful formal techniques that have emerged, the generative conception does not offer an appropriate explanatory model for human languages. A preferable alternative will be sketched, one that is based on a consideration of the ways in which a distinction can be drawn between those individual expressions that are well-formed (grammatical) and those that are ill-formed (ungrammatical).

The alternative is based on model theory rather than proof theory. In this approach, grammars are taken to be theories in the logician's sense—sets of statements in a logical description language—and the interpretation of that description language is given in terms of relational structures corresponding to the graphs that represent the structures of human language expressions. Languages and their related computational problems emerge looking rather different; for example, the recognition question, 'Is w one of the strings generated by the grammar G?', becomes 'Does any structure with the string yield w satisfy Γ?'; and the emptiness problem, 'Are any strings generated by G?', becomes 'Is Γ satisfiable?'.

It might be thought that the generative and model-theoretic approaches are equivalent ways of talking about exactly the same things. And indeed, with appropriate stipulations (specifically, restrictions on the class of candidate models), equivalence results can indeed be proved. If we permit the simplification of equating relational structures with the graphs that they represent, we can put it as follows. For certain classes of generative grammars $\mathcal{G}$ generating graphs in some class M we can find a description language $\mathcal{L}^M$ (interpreted on structures drawn from M) such that, for $X \subseteq M$, there is a $G \in \mathcal{G}$ such that $X = L(G)$ iff there is a set Γ of $\mathcal{L}^M$ formulas such that $X = \mathbf{Mod}(\Gamma)$. A celebrated early result of this sort obtained by Büchi [1] (and independently by Elgot [3] and Trakhtenbrot [7]) deals with the case where

$\mathcal{G}$ = regular grammars with some terminal vocabulary V_T
M = finite string models with points labeled from V_T
$\mathcal{L}^M$ = a weak monadic second-order (wMSO) language suited to M.

The result—now a staple of every course on finite model theory—is that wMSO on finite strings exactly characterizes the regular stringsets.

There are other much more recent results of this kind that have real importance, and through them new techniques have become available for establishing equivalences and inequivalences between frameworks for syntactic theorizing.

3. Although specific equivalences of the Büchi/Elgot/Trakhtenbrot sort can be established between descriptions in generative and model-theoretic modes, the light shed on linguistic phenomena is quite dramatically different under the generative and model-theoretic conceptions of grammars (see [5] for an earlier statement of this point of view). Among the most striking differences are the claims made about expressions that are partially but not fully well-formed; about the etiology of ill-formedness; about expressions containing undefined words; about the well-formedness of expression fragments; and about the possibility of clashes between constraints leaving gaps in the pattern of grammatical expressions.

One rather abstract issue relates to the claimed infinitude of expressions in human languages. It has been suggested that the expressions of a human language provide a rare (perhaps unique) case of countable infinity arising in a natural and entirely non-numerical domain. But the claim seems spurious, and certainly it is irrelevant to anything in linguistics [6]. Infinite cardinality is not a discovered property of human languages, and could not be. No good argument distinguishes the claim of infinitude either (countable or uncountable) from the claim of indefinitely large size, either with respect to the size of individual expressions or the size of the entire collection of grammatical expressions.

In this case, unfortunately, artifacts of generative theoretical techniques have been mistaken for properties of the phenomena. It is up to linguistic theorists to decide, on a case by case basis, whether some mathematical model of human language phenomena serves currently relevant research goals.

References

1. Büchi JR (1960) Weak second-order arithmetic and finite automata. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 6:66–92
2. Chomsky N (1959) On certain formal properties of grammars. *Information and Control* 2(2):137–167, reprinted in *Readings in Mathematical Psychology*, Volume II, ed. by R. Duncan Luce, Robert R. Bush, and Eugene Galanter, 125–155, New York: John Wiley & Sons, 1965
3. Elgot CC (1961) Decision problems of finite automata and related arithmetics. *Transactions of the American Mathematical Society* 98:21–51
4. Post E (1943) Formal reductions of the general combinatory decision problem. *American Journal of Mathematics* 65:197–215
5. Pullum GK, Scholz BC (2005) Contrasting applications of logic in natural language syntactic description. In: Hájek P, Valdés-Villanueva L, Westerståhl D (eds) *Proceedings of the 13th International Congress of Logic, Methodology and Philosophy of Science*, KCL Publications, London, pp 481–503
6. Pullum GK, Scholz BC (2010) Recursion and the infinitude claim. In: van der Hulst H (ed) *Recursion in Human Language*, no. 104 in Studies in Generative Grammar, Mouton de Gruyter, Berlin, pp 113–138
7. Trakhtenbrot BA (1962) Finite automata and monadic second order logic. *Sibirskii Matematicheskii Zhurnal* 3:101–131 (in Russian; English translation in *AMS Translations* 59:23–55, 1966)