[Mechanical Translation and Computational Linguistics, vol.9, no.1, March 1966]

The "Spectrum" of Weak Generative Powers of Grammars

by Wayne A. Lea, Research Laboratory of Electronics,
Massachusetts Institute of Technology, Cambridge

A *summary is presented of some results in the literature concerning the generative powers of various formal grammars. The relative generative powers are displayed graphically.*

I. Introduction

Many forms of grammars have been proposed in the study of such related language problems as mechanical translation, computer languages, mathematical linguistics, and the more general characterizations of natural languages. It is thus interesting to inquire about the relationships between such grammars. In particular, one might ask which proposed grammars are the "most powerful" (in some meaningful sense) and which are the most accurate characterizations of natural-language phenomena.

In this paper, grammars will be compared on the basis of the possible symbol sequences they may produce—that is, on the basis of what has been called their "weak generative powers." The relationships will be displayed on a "spectrum" of weak generative powers of grammars. It is hoped that this concise graphical display will be found an illuminating and useful comparative summary of grammars, generated languages, and equivalent machines.

No attempt will be made to explain in any detail the various grammars and machines listed in this paper, nor will the relationships discussed be proven, since they have already been considered in detail in various published papers. We shall merely consider a brief listing of each grammar, language, or machine type, and references where each relationship to other grammars, languages, and machines is shown. In listing references, our purpose is not to acknowledge the original developers of each interrelationship but, rather only to provide references where demonstrations of such relationships can be found. Although the author does not profess to have checked that all summarized results are valid, the literature indicates that they are. More important, the use of the chosen form of display clarifies any stated relationships between various formal grammars and proposed grammars of natural languages.

Thus, our goals are: (1) the listing of references where relationships between grammars, languages, and machines are presented and (2) the handy pictorial presentation (in a single "spectrum") of the relative weak generative powers of such grammars and their corresponding machines and resulting languages.

Though it is hoped that this listing and display of grammars will be in some sense exhaustive of known results, some possible grammar types may have been missed. One advantage of the spectrum display used herein (Fig. 1) is that such additions can be easily related to known grammars by simply marking them at the appropriate positions on the spectrum.

There are some known grammars whose relationships to other grammars are as yet unknown. The "branching" of the spectrum of Figure 1 will illustrate these uncertain relationships and thus indicate several unsolved problems in algebraic linguistics.

II. Languages, Grammars, and Machines

In combinatorial systems (see reference 1 or 2) and formal linguistic theory (reference 3, chap, iv), a *language* is simply a set of sequences or strings produced by concatenation of elements out of some finite vocabulary, set V_T. A *grammar G* is then a set of rules (or "productions") for enumerating the strings belonging to the language. A grammar may be precisely defined as a 4-tuple (V, V_A, S, P), where V is a finite non-empty *vocabulary,* V_A (called the auxiliary vocabulary) is a non-empty subset of V (and represents the symbols or phrase categories used at intermediate steps in the generation of a string), S (the axiom, or initial string) is a member of V_A, and P is a finite set of productions which yield strings in the *terminal vocabulary* ($V_T = V - V_A$) by substitutions starting with the axiom S. The *language* L *generated* by G is a subset of the free monoid V_T* generated by concatenating members of V_T. Terminal strings (members of L) are produced by *derivations* consisting of finite sequences of applications of the productions of G, starting from axiom S. A production of string ψ from string φ will be symbolized as $\varphi \rightarrow \psi$, while a derivation of ψ from φ is symbolized as $\varphi \Rightarrow \psi$.

To restrict the languages generated by grammars to interesting proper subsets of the free monoid V_T*, it is necessary to restrict the form of productions allowed.

* This paper is a revision of a memorandum written in June, 196.5, when the author was affiliated with the Mechanical Translation Group of the Research Laboratory of Electronics, Massachusetts Institute of Technology. The author acknowledges the co-operation and encouragement of several members of that group, including its director, Victor Yngve. The help of G. H. Matthews in providing references and reviewing early drafts of the paper is also acknowledged. This work was supported in part by the National Science Foundation (grant GN-244) and in part by the Joint Services Electronics Program under contract DA36-039-AMC-03200(E).

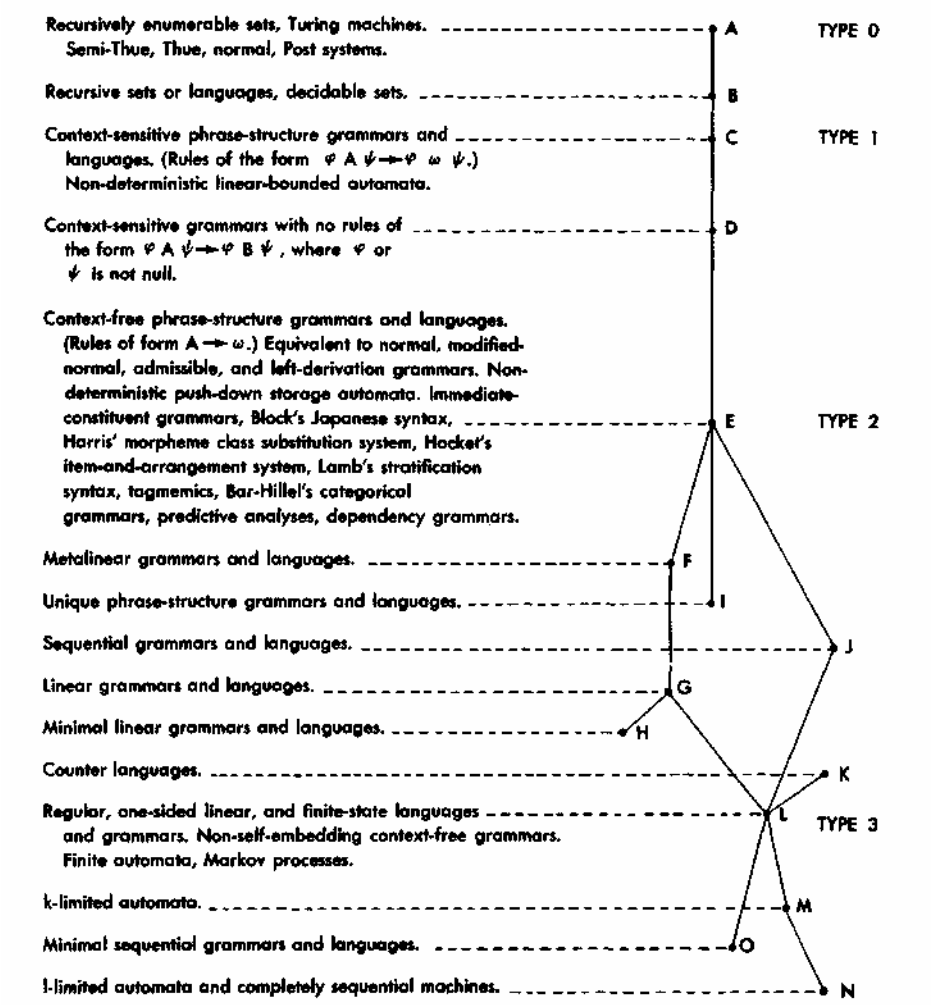

FIG. 1.—The "spectrum" of grammars, languages, and machines.

The broadest generative power of interest in mathematical linguistics is the power of a general Turing machine. Since Turing machines are associated with all effectively computable functions or algorithms,[3] broader generative power would involve sets which could not even be effectively (i.e., mechanically) enumerated.

III. The Spectrum

We shall now consider how the weak generative powers of various grammars and machines are related. Grammars are considered to be *weakly equivalent* when they produce the same language. *Types* of grammars are thus equivalent if for each language produced by a grammar of one type there is a grammar of the other type which produces the same language, and vice versa.

In accordance with the frequent use of *line diagrams* in set theory, whereby the inclusion of sets within others is pictorially displayed by showing successive subsets as successively lower points on a vertical line, the equivalences of grammars and the inclusion of certain languages within other types will be displayed as in Figure 1. The inclusion relation between languages is shown by the relative height on the line diagram or "spectrum"; points higher on the spectrum represent language types (sets) of which all lower points are special cases (subsets), resulting from added restrictions on the productions allowed in the grammars. Equivalent grammars are shown as a single point on the spectrum. (Thus, for example, the diagram illustrates the inclusion of all context-free languages within the set of context-sensitive languages, which are in turn included in all recursive sets, which are also in turn a proper subset of the recursively enumerable sets.)

The "branching" at the lower end of the spectrum indicates one of two types of relationship. Either it is not presently known how some such "branched" types of grammars, languages, or machines are related with respect to weak generative powers, or else the types are known to be incomparable with respect to inclusion. For example, it is not known whether all metalinear grammars are included within the sequential grammars, or vice versa, or whether they are inter-

secting sets, with some metalinear grammars not being sequential, and some sequential not metalinear. (Some of these questions may be easy to answer, but I have made no effort to do so. Perhaps the reader may attempt such studies.)

IV. References

The following is a list of references where each equivalence of grammars or machines is shown, or where certain grammars are shown to be properly included within other grammar types. The letter labeling each member of this list corresponds to the letter of the point on the spectrum which is presently being discussed.

A. Davis has shown (reference 3, chap. vi) the equivalences of *Turing machines, recursively-enumerable sets,* and combinatorial systems of *semi-Thue, Thue, Normal* and *Post* types. Chomsky (reference 1, theorem 2) has shown that his *Type 0 grammars* are equivalent to these systems. (The reader should be cautious when interpreting the present numbering scheme; Chomsky used a different one in reference 4).

B. In grammars, we may often be interested in determining whether or not a sentence is a member of a language set. Those sets for which this membership is effectively decidable are called *recursive* (or decidable) sets. Recursive or decidable sets are known to be a proper subset of recursively enumerable sets (see Davis, reference 3).

C. The productions used in semi-Thue systems may be restricted to those of the form $\varphi A\psi \rightarrow \varphi\omega\psi$, where a single symbol A is rewritten as a substring ω (non-null) and φ and ψ are strings from V*. This results in formal grammars called (after Chomsky) *context-sensitive phrase-structure grammars.* Chomsky has also called them Type 1 grammars and shown that the languages generated by such grammars are properly included in the set of recursive sets (reference 1, theorem 3). He also showed that such grammars are equivalent to grammars in which, for each rule $\varphi \rightarrow \psi$ the length of ψ is not smaller than that of φ.

Kuroda[5] has shown that a set is a context-sensitive language if and only if it is accepted by a non-deterministic linear-bounded automaton.

D. In reference 4 (pp. 365-67), Chomsky suggested that grammars with no rules of the form $\varphi A\psi \rightarrow \varphi B\psi$ (where A and B are single non-terminal symbols and either φ or ψ is not null) appear to be a proper subset of context-sensitive grammars and yet (as Parikh[6] had previously shown) contain context-free grammars (to be discussed under point E) as a proper subset.

E. When the rewriting of A as ω is unrestricted by the context φ—ψ, the context-free rules of the form $A \rightarrow \omega$ are obtained. Context-free grammars (with only rules of the form $A \rightarrow \omega$) have been shown to be a *proper* subset of context-sensitive phrase-structure grammars (Chomsky, reference 1, theorem 4). Context-free grammars are also called *Type 2 grammars.*

Context-free grammars have been shown to be weakly equivalent to *normal grammars* (which have rules of only the forms $A \rightarrow BC$ and $A \rightarrow a$, for $a \in V_T$, and thus represent binary trees[1,4]), *modified normal grammars* (with no pairs of rules A $\rightarrow BC$ and $D \rightarrow EB$ allowed), *admissible grammars* (in which every rule is used to generate *some* sentence and every generated string can be "completed" by further expansion into a terminal string, so no "dangling," unterminated derivations occur), and grammars with only *left derivations.* These facts are shown in references 1, 4, 7, and 8, respectively.

Gross[9] and Gaifman[10] have shown that dependency grammars are equivalent to context-free grammars.

It has also been shown that context-free languages are accepted by nondeterministic push-down storage automata.[4] Thus, a single point on the spectrum of weak generative power represents Type 2, or context-free grammars, normal grammars, modified normal grammars, admissible grammars, left- (or right-) derivation schemes, and non-deterministic push-down storage automata. Postal (reference 11, chap, iv; see also Chomsky, reference 12) has claimed that many grammars of natural languages, such as Block's Japanese syntax, Well's immediate-constituent grammars, Harris' morpheme class substitution system, Hockett's item-and-arrangement system, Lamb's stratificational syntax, and tagmemics all appear to be equivalent to context-free grammars. (Such demonstrations of equivalences as these between natural-language grammars and formal grammars depend, however, on the particular explicit, formal assumptions about the nature of vague, informal explications in natural-language descriptions. Thus, the formal assumptions often may be contested, with different assumptions implying different formal equivalences. For example, by suitably weak assumptions about stratificational grammars, they can be made to generate any recursively enumerable set, rather than just context-free languages. [I am indebted to Stanley Peters for this example.] The assumptions involved in the equivalences shown in Figure 1 are, however, apparently the prevalent ones in the literature.) Bar-Hillel's categorical grammars are shown to be equivalent to context-free grammars in reference 13. In reference 9, Gross shows a model based on *predicative analyses* to be equivalent to context-free grammars.

F. Chomsky and Schützenberger[14] have shown that the set of context-free languages properly includes the set of *metalinear languages.* Metalinear grammars have non-terminating rules of the form $A \rightarrow xBy$ or of the form $S \rightarrow \varphi$ and no rules of the form $A \rightarrow \varphi S\psi$ for any $A \in V$ and $\varphi, \psi \in V^*$.

G. Chomsky and Schützenberger[14] also showed that *linear grammars* (in which each non-terminating rule is of the form $A \rightarrow xBy$) are also a subset of metalinear

grammars, as is obvious from their form.

H. A proper subset of the linear languages is the *minimal linear* languages whose grammars have only one non-terminal (namely, the axiom) and rules of the forms $S \rightarrow xSy$ and $S \rightarrow c$, with the additional restriction that c does not appear in the x's and y's in the rules. Clearly, a minimal linear grammar is linear, but not all linear grammars are minimal.

I. *Unique phrase-structure grammars,* (which have rules of the forms $A \rightarrow x$, and $A \rightarrow yAz$, except for the axiom S, which introduces all non-terminals, including itself) are clearly a subset of context-free grammars, since each rule is a context-free rule. Apparently nothing else is known about their relative weak generative powers.

J. Ginsburg and Rice[15] have shown that all *sequential* grammars are context-free grammars and that they are *properly* included in the context-free ones. Sequential grammars are context-free grammars for which there exists an ordering of the non-terminal symbols such that for each i, j, if $A_i \Rightarrow \varphi A_j \psi$ then $j \geq i$ (or equivalently, ordered such that no rule $A_i \rightarrow \varphi A_j \psi$ for $j < i$). This restriction on the set of rules is such that if one symbol A_i is expanded into a string containing A_j, there is no derivation which, in turn, expands A_j into a string containing A_i.

K. *Counter languages* were discussed by Schützenberger in 1957 in an unpublished paper and, later, by Chomsky,[1] as being those produced by a device consisting of a finite automaton with an addition of a finite number of counters, each with an infinite number of positions. It is not known whether counter languages are all context free. But it is clear that the regular languages (to be discussed under point L) are all special cases of counter languages, with the number of counters equal to zero.

L. If all rules of a context-free grammar are restricted to the forms $A \rightarrow aB$ or $A \rightarrow a$, where $a \in V_T$ and $B \in V_A$, then what Chomsky[1] calls Type 3 grammars are obtained. Chomsky has shown (reference 1, theorem 6) that the languages produced by such grammars are exactly the *finite-state languages,* accepted (or produced) by finite-state automata (or Markov sources). Such languages are also referred to as *regular languages,* or *one-sided linear languages.* In reference 1 theorem 7, Chomsky showed that Type 3 languages are a proper subset of the Type 2 languages. Those Type 2 languages which are not Type 3 languages are necessarily *self-embedding* (that is, with derivations $A \Rightarrow \varphi A_j \psi$), according to Chomsky (reference 1, theorem 11), and what distinguishes Type 3 languages from arbitrary Type 2 languages is thus the lack of self-embedding.

All regular languages are found to make up a proper subset of the linear languages, as shown in reference 4, page 369.

Ginsburg and Rice[15] have shown that all regular or one-sided linear languages are properly contained within the set of sequential languages.

Chomsky[1] has shown (as was mentioned in point K) that all regular languages are special cases of counter languages.

M. A special type of automaton is a member of the set of "k-limited automata," whose state function is determined by the last k symbols of input sequence. Clearly, not every finite automaton is a k-limited automaton (reference 4, pp. 333-34.)

N. Those k-limited automata for which $k = 1$ are called by Ginsburg "*completely sequential machines.*"[16] Clearly, not every k-limited automaton is 1-limited.

O. A restriction on sequential grammars which does not allow recursive rules like $A_i \rightarrow \varphi A_i \psi$ gives *minimal sequential grammars.* It is apparent that minimal sequential grammars are all sequential, and their finite nature, due to not allowing reintroduction of symbols, makes them all regular, as well.

V. Relationship to Natural Languages

An interesting question relating to this spectrum of weak generative powers is how grammars of natural languages fit into the spectrum. That is, what are their apparent weak generative powers compared to those of the formal grammars discussed above? We have already seen that interest in being able to establish whether or not a string is a sentence of the language requires that the grammars be restricted to generative power less than or equal to that which generates the recursive sets. Furthermore, Chomsky has argued that the arbitrary permutations allowed by context-sensitive grammars are undesirable in grammars of natural languages (reference 1; see also reference 12 and reference 4, p. 365). Thus, powers less than those of arbitrary context-sensitive grammars seem to be needed for characterizing natural languages.

On the other end of the spectrum, it has been argued that natural languages can not be adequately generated by finite-state Markov processes. Furthermore, Chomsky and Postal have argued that there are many situations in natural languages where some context-sensitive rules are needed for adequate description, and thus generative powers greater than that of context-free grammars would appear to be required. These issues are discussed in references 1, 12, 2, and 11.

This, then, would result in the restriction of the range of weak generative powers for grammars of natural languages to a probable range between context-sensitive and context-free grammars, as is shown on the spectrum of Figure 1.

But at least one author would disagree with the above placement. In reference 17, the adequacy of a finite-state model is maintained.

The question of weak generative power is, of course,

only one factor in the determination of proper grammars of natural languages. Adequate structural descriptions of sentences and proper characterization of the interrelationships between sentences are additional factors to be considered.[2,4,11,12]

Received December 1, 1965

References

1. Chomsky, Noam. "On Certain Formal Properties of Grammars," *Information and Control,* Vol. 2 (1959), pp. 137-167.
2. Chomsky, Noam, and Miller, George A. "Introduction to the Formal Analysis of Natural Languages," in R. R. Bush, E. H. Galanter, and R. D. Luce (eds.). *Handbook of Mathematical Psychology,* Vol. 2, pp. 269-321. New York: John Wiley & Sons, 1963.
3. Davis, Martin. *Computability and Unsolvability.* New York: McGraw-Hill Book Co., 1958.
4. Chomsky, Noam. "Formal Properties of Grammars," in R. R. Bush, E. H. Galanter, and R. D. Luce (eds.). *Handbook of Mathematical Psychology,* Vol. 2, pp. 323-417. New York: John Wiley & Sons, 1963.
5. Kuroda, S. Y. "Classes of Lan-" guages and Linear-Bounded Automata," *Information and Control,* Vol. 7 (1964), pp. 207-223.
6. Parikh, R. "Language-Generating Devices," MIT Research Laboratory of Electronics, *Quarterly Progress Report No. 60,* Cambridge, January, 1961, pp. 199-212.
7. Greibach, S. "Inverses of Phrase Structure Generators." Ph.D. dissertation, Harvard University, June, 1963.
8. Matthews, G. H. "A Note on Asymmetry in Phrase Structure Grammars," *Information and Control,* Vol. 7 (1964), pp. 360-365.
9. Gross, Maurice. "On the Equivalence of Models of Language Used in the Fields of Mechanical Translation and Information Retrieval," *Information Storage and Retrieval,* Vol. 2, pp. 43-57. New York: Pergamon Press, 1964.
10. Gaifman, H. *Dependency Systems and Phrase Structure Systems.* (P. 2315.) Santa Monica, Calif.: RAND Corporation. 1961.
11. Postal, Paul. "Constituent Structure." (Publication 30.) Bloomington: Indiana University Center in Anthropolgy, Folklore, and Linguistics. (*International Journal of American Linguistics,* Vol. 30, No. 1 [January 1964]).
12. Chomsky, Noam. *Syntactic Structures.* The Hague: Mouton & Co., 1957.
13. Bar-Hillel, Y., Gaifman, C., and Shamir, E. *Bulletin of the Research Council of Israel,* Sec. F. Vol. 9, No. 1 (1960).
14. Chomsky, Noam, and Schützenberger, M. P. "The Algebraic Theory of Context-Free Languages," in P. Braffort and D. Hirschberg (eds.). *Computer Programming and Formal Systems,* pp. 118-159. Amsterdam: North-Holland Publishing Co., 1963.
15. Ginsburg, S., and Rice, G. H. "Two Families of Languages Related to ALGOL," *Journal of the Association for Computing Machinery,* Vol. 10 (1962), pp. 350-371.
16. Ginsburg, Seymour. *An Introduction to Mathematical Machine Theory.* Reading, Mass.: Addison-Wesley Publishing Co., 1962.
17. Yngve, V. H. "A Model and an Hypothesis for Language Structure," *Proceedings of the American Philosophical Society,* Vol. 104, No. 5 (October, 1960), pp. 444.466.
18. Chomsky, Noam, and Miller, George A. "Finitary Models of Language Users," in R. R. Bush, E. H. Galanter, and R. D. Luce eds.). *Handbook of Mathematical Psychology,* Vol. 2, pp. 419-491. New York: John Wiley & Sons, 1963.
19. Landweber, P. S. "Three Theories on Phrase Structure Grammars of Type 1," *Information and Control,* Vol. 6 (1963).
20. McNaughton, R. "The Theory of Automata," *Advances in Computers,* Vol. 2. New York: Academic Press, 1961.
21. Matthews, G. H. "Discontinuities and Asymmetry in Phrase Structure Grammars," *Information and Control,* Vol. 6 (1963), pp. 137-146.