The generative power of rule orderings in formal grammars*

FRANCIS JEFFRY PELLETIER

Abstract

One way of 'restricting linguistic theory' is the L-view: place sufficient restrictions on the allowable rules of grammars so as to reduce their generative power. Another way is the G-view: disallow certain grammars, regardless of whether this results in a reduction of generative capacity. The present paper adopts the L-view and, consequently, investigates the generative power of various theories.

One area in linguistics where restrictions on linguistic theory have been advocated is in the ordering (within the cycle) of the application of the rules which generate the language. We consider eight proposals: Total Ordering; Partial Ordering (= Total Ordering plus iterative application); Semi Ordering (= Anderson's 'local ordering' without iterative application); Semi Partial Ordering (= Semi Ordering plus iterative application); Unorderings (= Ringen 'Condition VI, *unmodified'); Quasi Orderings (= Ringen 'Condition* VI, *modified'); Random Orderings; and Simultaneous Application.*

If, for any grammar obeying rule ordering conditions A there is a grammar obeying rule ordering conditions B which contains exactly the same class of derivations, then rule ordering theory B is at least as powerful in strong generative capacity *as rule ordering theory A. Similar considerations are used to define the notions of* equivalent, more powerful, *and* non-comparable *in strong generative capacity. A series of theorems are proved showing the relative strength of the eight rule ordering theories.*

Some linguists who advocate 'random ordering' actually have in mind random ordering plus some 'universal principles'. We investigate the effect of four of these principles from the standpoint of the L-view, showing that two of them are strongly equivalent to total orderings and that two of them are intermediate between total and partial orderings.

We close with an indication of what the role of mathematical linguistics should be for the ordinary working linguist.

Linguistics 18 (1980), 017–072.

0024–3949/80/0018–0017 $2.00

1. Introduction

This paper is an essay in the philosophy of linguistics. It combines elements from the study of formal grammars (standard references are Hopcroft and Ullman, 1968; Gross and Lentin, 1970) with elements from standard linguistic theory (particularly Chomsky, 1965, as formalized by Peters and Ritchie, 1973a; or Ginsburg and Partee, 1969), and elements from recent discussions of the effects of rule orderings in linguistics (Pullum, 1979, is a definitive source, we shall also discuss such authors as Ringen, 1972; Levine, 1976; and Koutsoudas *et al.*, 1974), and finally with more theoretical discussions of such issues as simplicity of a (linguistic) theory, the possibility of language learning, and so on (we shall look at Derwing, 1973; Chomsky, 1971, 1975, 1977; Lasnik and Kupin, 1977; and Wasow, 1978, among others).

We start (Section 2) with a very brief explanation of the major methods and results of the study of formal grammars, and follow this with a brief discussion of "standard" linguistic theory (Section 3). Readers familiar with one (or both) of these areas can skip it (them) and go directly to Section 4, which is a discussion of the notions of grammar, theory of grammar, and generative power of a grammar or theory of grammar. This is followed by a statement of various results known about the generative power of linguistic grammatical theories, and then (Section 5) by an outline of the precise topic of this paper and a survey of what others have said about it. Section 6 is devoted to defining some requisite preliminary notions before we present (Section 7) a series of 'theorems' concerning the generative power of various kinds of rule orderings. We then discuss (Section 8) some actual linguistic proposals that have been made about 'conditions upon application of rules' as a replacement for ordering rules. Various theorems are proved to show that these 'conditions' have no real empirical consequences. We close (Section 9) with a suggestion about what the role of mathematical linguistics should be for linguists.

2. Formal grammars

One way to study a language is to view it as a set of strings (consisting of words or of letters, etc.) which is produced in accordance with a set of rules. The strings are said to be composed of *terminal symbols*, which comprises the *alphabet* or *terminal vocabulary* and is denoted Vt. A *sentence* then is a concatenation of members of Vt. We shall let *e* denote the 'empty string'; $^{+}$Vt is the set of all possible strings not containing *e*, and *Vt is the set of all possible strings (which may include *e*).

The sentences of languages (at least of natural languages) also have structure, in addition to being merely members of $^{+}$Vt or *Vt. In natural languages, we typically call some subsequences of $^{+}$Vt 'noun phrases' (NPs) or 'verb phrases' (VPs), for example; and we traditionally call an entire terminal string a 'sentence' (S). Such symbols are part of our *non-terminal vocabulary*, Vn. The entire vocabulary we denote by V, that is the union of Vn and Vt (and we add the requirement that Vn and Vt are disjoint); *V is the set of all strings of V (including *e*).

The *grammar* of a language is a way of organizing all of the potentially infinite members of *V into some finite mode of representation. Toward this end, we introduce the set of *rules* of the grammar, which we denote by ($\rightarrow$). The relation ($\rightarrow$) holds between members of *V; if $\langle \alpha, \beta \rangle \in (\rightarrow)$, we shall write $\alpha \rightarrow \beta$. We shall use S as our 'initial symbol'. A *sentence generated by the grammar G*, then, will be an $x \in$ *Vt such that there is a sequence $S \rightarrow \alpha$, $\alpha \rightarrow \beta, \ldots, \gamma \rightarrow x$. The *language generated by G* is the set of all sentences generated by G.

It is quite easy to show that placing certain kinds of restrictions on the ($\rightarrow$) relation will result in different languages. For example, if every member of ($\rightarrow$) has the property that it is of one of the two forms

$A \rightarrow xB$
$A \rightarrow y$

where A, B $\in$ Vn and x, y $\in$ $^{+}$Vt, then the language being generated will be a Kleene-regular set. It can furthermore be shown that every Kleene-regular set is generated by some such grammar (called 'right linear grammars' or 'type 3 grammars').

If the rules are required to be of the form

$A \rightarrow B$

where A $\in$ Vn and $B \neq e$, then the grammar is 'context free' or alternatively a 'type 2' grammar.[1] If the rules are required to be of the form

$\alpha A \beta \rightarrow \alpha B \beta$

where $\alpha, \beta \in$ *V and A $\in$ Vn and $B \neq e$, then the grammar is 'context sensitive' or a 'type 1' grammar. It is common practice in linguistics to write these context sensitive rules in this way:

$A \rightarrow B \quad / \alpha --- \beta$

If the rules have no restrictions on them, the grammar is said to be a 'unrestricted rewrite system' or a 'type 0' grammar. It is straightforward to prove that type 0 grammars give the recursively enumerable (r.e.) sets. Early results of Chomsky's show that L_3 (the class of languages that can be generated by a type 3 grammar) is properly included in L_2 which is properly included in L_1 which is properly included in L_0 (see Hopcroft and Ullman, 1968).

Intuitively speaking, a r.e. set is any set producible by some rule(s) or other. If one adds the further requirement that one is able to determine, in some finite time, whether an arbitrary string is or isn't in the set, then the set is said to be recursive. The r.e. sets properly include the recursive sets, which properly include those produced by context-free grammars. We shall have occasion to return to recursive sets (and the grammars that produce them) below.

We can mirror the structure of terminal strings by adding a convention so that each rule introduces an appropriate pair of brackets corresponding to non-terminal symbols altered or rewritten in the derivation of a terminal string by the rules of the grammar. Thus, for a rule like:

$$\alpha A \beta \rightarrow \alpha B \beta$$

for example, the convention would have the effect of surrounding B by brackets labelled A. Alternatively, we could write this directly into the rule:

$$\alpha A \beta \rightarrow \alpha [_{A}\ B\ ]_{A} \beta$$

It is then trivial to induce a 'bracketization' function on terminal strings which will yield their (surface-) structural descriptions, and a 'de-bracketization' function which will yield the terminal string without the brackets. This is in fact the method proposed in Peters and Ritchie (1973a), in Langendoen (1979), and in Crespi-Reghizzi (1971). When looked at this way, it is tempting to say that the labelled brackets are part of the terminal string itself, that is, part of the language being generated. In the sections that follow, we shall in fact view the matter in precisely this way.

It should perhaps be mentioned that the interpretation of formal grammars presented here is the 'orthodox interpretation', where rules like A → B C are string-mapping rules: it maps a string containing an occurrence of the symbol A into a string containing the substring BC in place of the occurrence of A. However, rules can be viewed otherwise. McCawley (1968) proposed that one look at rules as 'node admissibility conditions' which admit fragments of a parse tree. That is, they tell one whether a given parse tree is legitimate. Besides McCawley's remark that this is more true to linguistic practice, various formal benefits have been claimed to follow from this way of looking at rules. It has been proved in Peters and Ritchie (1973c), Joshi and Levy (1977), and Gazdar (1979a) that, under various different interpretations of 'analyze' and 'parse', a language is context free if and only if there is a finite set of context sensitive rules that parse the language; i.e. if and only if there is a collection of trees whose terminal symbols are the sentences of the language and a finite set of context sensitive(!) rules which exactly analyze these trees. Thus node admissibility rules can be stated as context sensitive, but nonetheless the language 'generated' (= analyzed) will be only context free. Furthermore, as Gazdar has pointed out (personal communication), such an interpretation obviates any question about rule ordering, since rules cease to be the kind of things that can be ordered. In particular, this obviates the difficulties discussed below (Section 6) concerning the possibility of 'simultaneous application' of rules that dictate 'contradictory changes'. Similar remarks from an informal point of view can be found in Derwing (1973). Peters and Ritchie (1973c:333) claim:

> ... all sets of context-sensitive rules which have been used as a part of transformational grammars describe the same set of labeled bracketings whether interpreted as rewriting rules or immediate constituent analysis rules. These rule sets therefore describe only context-free languages. This means that the power to describe non-contextfree languages has not been used by linguists, although it is implicit in the use of context-sensitive rules as rewriting rules.

Gazdar (personal communication) urges this as a clear case where eliminating the possibility of rule ordering from a type of grammar (viz., context sensitive grammars) results in a constraint on weak generative capacity. This is relevant to the general issue to be developed in this paper: the effect of rule orderings on the generative power of grammatical theories. But nonetheless it will not be considered, and in fact this paper is exclusively concerned with the 'orthodox interpretation' of rules as string-mappings. The reasons for this are mainly ones concerning the fact that the 'orthodox interpretation' is standard in linguistic theory, and the discussions in the literature concerning rule ordering presuppose it.

3. 'Standard' linguistic theory

'Standard' linguistic theory is whatever is proposed in Chomsky (1965) for syntax and Chomsky and Halle (1968) for phonology. This is, to be sure, not current linguistic theory; but it does contain all the resources necessary to reconstruct current theory.[2] We concentrate here on explaining syntax. In 'standard' linguistic theory, grammars of natural languages are said to be structured in the following way.

(1) There is a set of Phrase Structure Rules (PS rules), with initial symbol *S*. Such rules might look like the following, where the parenthe-

sized elements amount to an abbreviatory convention which 'collapses' two rules — one which has the unparenthesized element in it and one which has no such element in it at all. (In linguistic terminology such parenthesized elements are said to be 'optional'.)

S → NP Aux VP
NP → (Det) N Num (S)
N → CS
.... etc ....

These rules have the effect of yielding a set of sentences (and their structure) which look like:

$$[_{S}\,[_{NP}\,[_{Det}\;]\,[_{N}\,CS]\,[CS]\,[_{Num}\,Pl]]\,[_{Aux}\ldots$$

(As we said in Section 2, we can treat the labelled bracketing as a part of the terminal vocabulary, so that the above string is composed entirely of terminal symbols; but we can also induce the *B* ('bracketing') function on it so as to recover just the structure if we so desire). When looked at this way, PS rules form a context-free grammar with initial symbol *S*. It is common in linguistics to represent the bracketing by the equivalent means of labelled trees.

(2) There is a Lexicon which contains elements that are ordered pairs. The first member of these ordered pairs is an 'underlying phonological shape' of some morpheme of the natural language. This is generally given by stating the 'non-redundant distinctive features' of the morpheme, e.g. we might have:

$$\begin{bmatrix} +\text{nasal} \\ +\text{coronal} \end{bmatrix} \quad \begin{bmatrix} +\text{vocalic} \\ +\text{coronal} \end{bmatrix} \quad \begin{bmatrix} +\text{strident} \\ +\text{apical} \end{bmatrix}$$

The second member of the ordered pair is a statement of a 'context', for instance:

$$\begin{bmatrix} ([+\text{Det}])\underline{\qquad}[+\text{VP}] \\ +\text{N} \\ +\text{Common} \\ \vdots \end{bmatrix}$$

The relevant rule regarding 'lexical insertion' is that the symbol *CS* in the terminal strings of the PS rules may be replaced by the first member of the ordered pair if the context of this terminal string does not violate the context as stated by the second member of the ordered pair. It will be noted that, when done this way, such 'lexical insertion' is a kind of context-sensitive rule. The language which is the output of this lexical insertion into the terminal strings of the PS rules is often called the 'deep structure' or the 'base language' of the natural language which is to be generated.

(3) Given the base language as input, transformational rules alter the structure and linear order of the terminal strings of the base language. For example, there might be a rule ('there-insertion') which does the following (where *W*, *X*, and *Y* are variables over (possibly empty) strings).

$$[_{S}[_{NP}[_{\substack{N\\ -\mathrm{Def}}}\ W]]X\ \textit{be}\ Y \rightarrow [_{S}\,\textit{there}\ X\ \textit{be} + [_{NP}[_{\substack{N\\ -\mathrm{Def}}}\ W]]\ Y]$$

which would alter sentences like:

$$[_{S}[_{NP}[_{\substack{N\\ -\mathrm{Def}}}\ \text{a boy}]][_{VP}\text{past-be}[_{PP}\text{on}[_{NP}\text{the dock}]]]]$$

to

$$[_{S}\text{there}[_{VP}\text{past-be} + [_{NP}[_{\substack{N\\ -\mathrm{Def}}}\ \text{a boy}]][_{PP}\text{on}[_{NP}\text{the dock}]]]]$$

(i.e. alters *A boy was on the dock* to *there was a boy on the dock* and suitably alters structure).

These transformational rules apply *cyclically*: they start with a most deeply embedded *S* and work outwards, only considering the structure *within* the particular *S* part of the entire string that is being generated.[3] Within any given cycle there are sufficient restrictions on deletion that transformational rules are not unrestricted re-write rules (see Peters and Ritchie, 1973b), but with the addition of the principle of cyclic application they are equivalent to Turing machine rules (see Peters and Ritchie, 1973a). This is discussed in more detail in the next section.

(4) The Phonological Rules apply to the terminal symbols of the output of the transformational rules. That is, they apply to the symbols which were introduced by the lexical insertion rules, and they convert them into a representation of a pronounceable word of the natural language. The usual formulation of phonological rules makes them be context-sensitive rules.

4. Grammars, theories of grammar, generative capacity and constraints

A *grammar* is a particular set of rules. A *theory of grammar* is a set of principles that picks out a certain group of grammars as being legitimate grammars. Thus, for example, the sample grammars mentioned in Section

2 are each grammars. The context-free theory of grammars picks some of them out as being legitimate (legitimate context free, that is). More generally, a theory of grammar is a statement of what is in principle permissible in a grammar.

A grammar produces (or generates) a language (set of terminal strings) from a given input. A theory of grammar produces a *set* of languages from a certain *kind* of input. As indicated in Section 2, some theories of grammar allow this set of languages to be characterized in alternative ways (e.g. 'all regular expressions', 'all recursively enumerable sets'). That is, the terminal strings of any grammar of the theory also obey some alternative characterization. The ability of a theory of grammar to produce a certain class of languages (terminal strings) from a given kind of input is called its *weak generative capacity*. Traditional studies in formal grammar and in the formal properties of linguistic theories of grammars have concentrated exclusively on the weak generative capacity of the grammatical theories, i.e. on the sets of terminal strings produced by the grammars, and have not consciously investigated the linguistically more interesting issue of the 'structural descriptions' (= labelled bracketing) of the terminal strings.[4] The ability of a theory of grammar to produce a certain class of languages and associated structural descriptions from a given kind of input is traditionally called its *strong generative capacity*. The distinction between strong and weak generative capacity of a grammar has been characterized by Chomsky (1965:60):

> Given a theory of language structure, we can distinguish *weak generative capacity* from its *strong generative capacity* in the following way. Let us say that a grammar *weakly generates* a set of sentences and that it strongly generates a set of structural descriptions (recall that each structural description uniquely specifies a sentence, but not necessarily conversely) .... The study of strong generative capacity is related to the study of descriptive adequacy ... A grammar is descriptively adequate if it strongly generates the correct set of structural descriptions. A theory [of grammar] is descriptively adequate if its strong generative capacity includes the system of structural descriptions for each natural language ...

It is clear from Chomsky's discussion that a descriptively adequate grammar has to assign the appropriate constituent structure to strings of terminal symbols. (Appropriate in the sense of being in accord with native speakers' intuitions, at least when they have any). But as we indicated above, both in Section 2 and Section 3, such a desideratum can be accomplished within the output language itself when augmented by our bracketization function. Thus for example, the allegedly ambiguous string:

(1) We are visiting scholars

is to be distinguished on the grounds that our grammar outputs both:

(2) $[_{S}\,[_{NP}\text{We}]\,[_{Aux}\text{are}]\,[_{NP}\,[_{Adj}\text{visiting}]\,[_{N}\text{scholars}]]]$

(3) $[_{S}\,[_{NP}\text{We}]\,[_{VP}\,[_{Aux}\text{are}[_{VP}\,[_{V}\text{visiting}]\,[_{NP}\,[_{N}\text{scholars}]]]]$

(These structures are not to be taken too seriously, we use them only as examples. The point is that this kind of ambiguity *could* be handled as being different terminal strings of our grammar, and that the bracketization function would associate two distinct structural descriptions for (1). And if it is treated in this manner, the detection of such ambiguity should be called part of the weak generative capacity of the theory of grammar.) Chomsky (1965:89–90) approves of this method of understanding rules, although he mysteriously claims that the addition of such an understanding of the PS rules no longer constitutes a PS grammar, even though the two are equivalent in weak generative capacity.

However, the surface-structural ambiguity illustrated by (1) is not the only kind of ambiguity countenenced in linguistics.[5] Sentence (4) seems to have only one bracketization, namely something like (5), even though it (arguably) is *syntactically* ambiguous; the ambiguity having to do with whether the *when* refers to the time Mary said something or the time of Mary's departure.[6]

(4) When did Mary say she would go?

(5) $[_{S}\,[_{Adv}\text{When}]\,[_{Aux}\text{did}]\,[_{NP}\text{Mary}]\,[_{VP}\,[_{V}\text{say}]$
$[_{S}\,[_{NP}\text{she}]\,[_{VP}\,[_{Aux}\text{would}]\,[_{V}\text{go}]]]]]$

It is common practice to regard this ambiguity as arising from 'different deep structure, same surface structure'. This too (the ability to make such assertions) is apparently to be called part of the strong generative capacity of a theory of grammar. As Chomsky says (1965:140–141):

> ... the grammar defines the relation "the deep structure M underlies the well-formed surface structure M' of the sentence S" and, derivatively, it defines the notions "M is a deep structure," "M′ is a well-formed surface structure", "S is a well-formed sentence", and many others (such as "S is structurally ambiguous", "S and S′ are paraphrases", "S is a deviant sentence formed by violating rule R or condition C").

So it would seem that in addition to including facts about the surface structure, the concept of strong generative capacity is to include *all* the facets relevant to determining whether a grammar is descriptively adequate, even those which invoke the notion of a derivational history. Similarly, Wall's (1972:290) interpretation of 'strong generative capacity' seems to include derivational history:

> ... nearly all the work on formal grammars deals exclusively with the sets of strings they generate (called the *weak generative capacity* of the grammar) and little has been said about the kinds of structural descriptions (constituent-structure trees) assigned to the grammatical strings (*strong generative capacity*). A natural language grammar must, of course, not only generate the correct set of strings but it must also specify correct structural descriptions — "correct" in the sense that they agree with the speaker's intuitions in marking a sentence as n-ways ambiguous, marking two sentences as paraphrases of each other, specifying the grammatical relations that hold between parts of a sentence etc.

However, we diverge from Chomsky and Wall in that the structural descriptions of a sentence that we can induce by the bracketization function will be considered by us to be part of the weak generative capacity of a grammar. Since weak generative capacity is properly included in strong generative capacity, this makes no real difference when one is investigating strong generative capacity.[7] But we do wish to emphasize the sometimes-neglected point that some overall properties of a grammar cannot be included in its weak generative capacity (such as relations between deep structure and surface structure), and any such properties shall be treated here as part of the grammar's strong generative capacity. Even though the point is sometimes neglected, it clearly is nascent in both the Chomsky and Wall quotes just given, for some of the empirical relations mentioned by them in describing strong generative capacity are defined in terms of derivational history of sentences. Paraphrase is: derived from same deep structure. Ambiguity is: same surface structure derived from distinct deep structures. Grammatical relations holding between parts of a sentence is: being related in specified ways to the same underlying segment. Thus an investigation of strong generative capacity will be concerned with the derivational histories of the output sentences.[8] More generally, it is said that a comparison of two theories of grammar in terms of their weak generative capacity has only one empirical reflex: grammaticality. In other words, it amounts to asking the question, if one theory can call a certain (kind of) sentence grammatical (= generate it with some grammar obeying the theory) can the other? Even with our strengthened notion of weak generative capacity, grammaticality only amounts to: can the theories generate it and its structural description? Strong generative capacity is supposed to compare theories of grammar in respect to *all* their empirical reflexes: in addition to grammaticality, can the two theories give identical paraphrase claims? Can they give identical ambiguity claims? Can they relate the structural descriptions of the members of the output language to the intermediate stages of derivation in the same way?, etc. Strong generative capacity is not a totally defined notion — there are too many things someone *might* want to call 'an empirical reflex'. We here shall consider any feature of the surface structure of a sentence and any feature of the derivational history of a sentence to be included as an empirical reflex; and we consider all such features to be under discussion when we talk about the strong generative capacity of a theory of grammar. If two theories of grammar are equivalent in strong generative capacity, there is no way to empirically distinguish them; they are, in the strictest and most pure sense, notational variants of one another.[9]

Bearing in mind what we include in weak generative capacity versus strong generative capacity, we shall define various *relative* notions of 'strength of a theory of grammar'. One *theory* of grammar (i.e. a statement of what is in principle permissible in a grammar subsumed under the theory) is *at least as strong in weak generative capacity* as another if, given an input, every language produced by some grammar obeying the strictures of the latter can also be produced by some grammar obeying the strictures of the former. If the converse can also be established, the two theories of grammar are *equivalent in weak generative capacity.* If one direction can be established, and they are not equivalent, then the first is *stronger* (*or: more powerful*) *in weak generative capacity* then the second. If they each can be shown to be able to produce languages the other cannot, then they are *non-comparable in weak generative capacity.* Substituting 'strong' for 'weak' and 'empirical reflex' for 'language' in the preceding definitions gives us analogous characterizations of the notion of relative *strong* generative capacity of theories of grammar.

Peters and Ritchie (1973a: Section 2) define a transformation as the mapping induced on an n-term structural condition by an n-term transformational mapping. The transformational mapping is a composition of four elementary rules (deletion, substitution, left-adjunction, and right-adjuction) and must satisfy 'the principle of recoverability of deletions' (roughly: one can delete some non-terminal of a string if there is a copy of it somewhere still left in the structural change, and one can delete a specified terminal symbol). A transformational grammar *G* is a pair $\langle L_1, T\rangle$ where L_1 is a base or input language (e.g. a PS grammar's output) and T is a sequence of transformations. The language, L(G), output by the transformational grammar is the set of (labelled) strings of terminal symbols generated in accordance with a cycling principle (see their Definitions 4.1, 4.2, 4.3). Applying the 'de-bracketization' function to these labelled strings in order to get the unlabelled strings (call these

L′(G)), one can prove the following theorem remarkably easily:[10]

(Peters and Ritchie 5.1): Every recursively enumerable language is the language L′(G) of some transformational grammar G with a context-sensitive base language.

The proof of 5.1 procedes by noting that an arbritrary r.e. language *R* can be generated by a grammar with all rules of the form a*X*b → a*Y*b (*X* and *Y* elements of *Vn) or A → *a* (A element of Vn, *a* of Vt). So take this grammar, and for every rule A → B in which B is *n* elements shorter than A, replace it by the rule A → BC(n). (C is a new non-terminal, here repeated *n* times). Add these two sorts of rules: CB → BC, for all the old non-terminals B, and C → *c* (*c* a new terminal symbol). First note that this new grammar G′ is a context sensitive grammar. Note next that this new language R′ generated by this grammar is identical to R except for occurrences of *c* as rightmost symbols of a cyclic node. The principle of recoverability of deletions does not prevent the addition of a rule to delete this terminal symbol (since we delete specified terminal symbols in specified contexts). Thus for any r.e. language R, there is a context sensitive base language G′ (the rules we constructed) and a transformational component (the deletion transformation) which generates R.[11] Peters and Ritchie (1971) also show that there is a context-free base language which will generate R, but that such a proof has the transformational grammar make heavy use of 'filter functions'.

These results are widely thought to show there to be something drastically amiss in linguistics. First, during the period when the 'standard' theory was in vogue, it was hoped that linguistics could find a Universal Base Language — a deep structure language that every natural language shared, and to which one merely added (1) a lexicon appropriate to a given language, and (2) some particular subset of all the permissible transformational rules, in order to generate the given language. As Peters (1970) points out, the hope of finding such a universal base language is too easy to satisfy, since transformations could be found for any proposed base which will do the trick. The hope for a universal base language has not been prominent in recent discussions, possibly due to this result. The second fault with 'standard' linguistic theory was thought to be that if linguistic theory could describe or generate *any* r.e. language, then it was too powerful a theory for describing the (supposed) underlying mental reality of learners (speakers) of the language. For, it was said, a child learns language on the basis of a small amount of data and in a short period of time; hence, the mental realities underlying a child's linguistic ability must be tightly constrained so that he comes up with the correct grammar easily and quickly. But doesn't this result show, it was asked, that this underlying mental reality isn't correctly characterized by a transformational grammar? For, if it were, the child could never learn the proper grammar, there being simply too many of them which would do the job. Furthermore, it was alleged,[12] the fact that people can classify arbitrary strings as being grammatical or ungrammatical shows that natural languages must be recursive sets, and not arbitrary r.e. sets. Thus transformational grammar is too powerful a theory, and must be restricted in some way to generate only recursive sets.

How is a theory of grammar to be restricted or constrained? In a very interesting paper, Wasow (1978) considered two ways of applying constraints on a theory. The direct way, which Wasow calls the 'L-view', is to place sufficient constraints on the theory so that the class of languages generated by the grammar is smaller. In the terminology of above, this would be to alter the grammars allowed so that the weak generative power of the theory is lessened. An alternative view, which Wasow calls the 'G-view', would be to place restrictions on the theory which will limit the kinds of grammars thereby allowed, regardless of whether such limitations affect the weak generative capacity of the theory. As Wasow points out, these are not equivalent types of limitations. A restriction on a theory of grammar to the effect that no terminal symbol could be immediately dominated by a branching node would restrict the permissible grammars that theory allows. But since a grammar that obeys this restriction (but is otherwise identical to one that doesn't) could be trivially defined from any grammar which doesn't, it follows that the class of languages generated by these two theories of grammar is identical. Of course limiting the class of languages produced entails a limitation on the forms of grammars, but as this example shows, the converse is not true.

According to the G-view, the point of putting constraints on theories of grammars is not to limit the class of languages that can in principle be generated by some grammar obeying the theory, but rather to constrain the kinds of descriptions available to the linguist in describing the language. And the reason for desiring this kind of constraint is the hope that such constraints will sufficiently circumscribe what it is that a speaker has learned so that it is reasonable to claim that a child could learn *that*. This point has been put by Lasnik and Kupin (1977):

> We follow Chomsky (1965) in the belief that children acquire their grammar from an environment that seriously underdetermines it, and that some evaluation metric is employed to select the appropriate grammar for any particular language. Certainly if the class of possible grammars is smaller, the evaluation task becomes simpler. By restricting the class of allowable grammars, we thus approach an explanation of how language can be acquired. (p. 174).

They then develop such a restricted theory of grammar and conclude with the following remarks:

> It seems clear to us that our theory shares the defect of the *Aspects* [Chomsky 1965] theory noted by Peters and Ritchie [1973a] .... our theory provides a grammar for every r.e. set .... In comparing two theories, it is reasonable to abstract away from their common virtues and shortcomings. In the present instance, such an abstraction leaves our theory much less powerful .... Notice that we use the term "powerful" not with respect to the character of the languages generated but rather with respect to the relative size of the classes of grammars allowed. (p. 195).

Chomsky (1965:62) puts the G-view as follows:

> It is important to keep the requirements of explanatory adequacy and feasibility in mind when weak and strong generative capacities of theories are studied as mathematical questions. Thus one can construct hierarchies of grammatical theories in terms of weak and strong generative capacity, but it is important to bear in mind that these hierarchies do *not* necessarily correspond to what is probably the empirically most significant dimension of increasing power of linguistic theory ... Along this empirically significant dimension, we should like to accept the least "powerful" theory that is empirically adequate. It might conceivably turn out that this theory is extremely powerful (perhaps even universal, that is, equivalent in generative capacity to the theory of Turing machines) along the dimension of weak generative capacity, and even along the dimension of strong generative capacity. It will not necessarily follow that it is very powerful (and hence to be discounted) in the dimension which is ultimately of real empirical significance.

The G-view is thus making a certain kind of claim about 'psychological reality': The constraints put on a theory of grammar are, in some intimate sense, mirrors of the actual states of mind of a speaker of the language. One is tempted to ask whether the methodology embraced by the G-view is sufficient for the task it sets itself. Let us briefly indicate why it seems to hold little promise in the form in which it is actually practiced.[13] In practice, the methodology of the G-view runs like this: A corpus of linguistic data can be 'economically' and 'revealingly' described by using a certain descriptive device *D* than it can without using it. (Such a device might be some abbreviatory convention, or some ordering statement on rules, or a more esoteric constraint like 'the heavy NP constraint', 'the strict cycle condition', 'the structure preserving constraint', 'the A-over-A principle', etc.) The claim is then made that the 'psychological, internalized grammar' of language users embodies this constraint. It is usually added that this is an *empirical* hypothesis — speakers do not *have* to use this constraint, but they in fact do. Of course, the fact that data can be 'economically' and 'revealingly' described by the theory does not by itself support the claim about psychological reality; to argue in this manner would be the most blatent circularity imaginable. We therefore wish to have some other method of checking on the predictions made by these constraints.[14] But it is here, we think, that insuperable difficulties attend the G-view; for the G-view must now define some measure which maps their proposed constraint onto speaker psychology, and then a function which maps this proposed speaker psychology onto some observable correlate (since it is, after all, to be an *empirical* hypothesis). Straightforward attempts to do this (such as: transformationally-defined complexity is mapped onto length of time to understand, or ease of making errors, or difficulty in memorizing, or time of child-acquisition, or slips of the tongue, or aphasic difficulties, etc.) have all been shown not to be well-correlated.[15] The standard reply to this lack of straightforward correlation is to reply that other factors 'of a performance nature' (short-term memory limitations, etc.) interfere with the predictions of the proposed grammar, which (it is claimed) is supposed to describe the speaker's *competence*, not his performance. But this reply simply reinforces the claim that the G-view can have no empirical support until these psychological or 'performance' factors are first brought under some theory so that one knows what the function mapping the linguistic theory onto the observable correlate is. None of this, of course, is to say that the proposed constraints are not in principle empirical; it is rather to say that it is premature to propose such constraints without there being any hope of evaluating them in the foreseeable future.

Transformational grammarians have, for the most part, been advocates of the G-view. This is (no doubt) in part because Chomsky, still the leading linguist in the world, is an exponent of the G-view. However, agreeing with Wasow (1978), we think a better approach is embodied in the L-view. The view of constraints here is to limit the class of possible human languages, thereby making claims about the limitations of human language acquisition. As Wasow points out (1978:85), such limitations make predictions (at least about boundary conditions) on such notions as parsing time and other processing features, since the exclusion of certain classes of languages also excludes certain classes of grammars.

There have always been exponents of the L-view since the beginnings of transformational grammar. Putnam (1961) made various suggestions about constraining the theory of transformational grammar so that the class of languages it generates is some subset of the r.e. languages. One method, he notes, would be to demand that no transformational rule ever yield an output shorter than its input. Another method would be to establish two upper limits x and v on any transformational grammar so that at most x terminal symbols can be deleted by any one transformation and at most v such transformations can be applied in any derivation. Such restrictions amount to constraining the class of languages generated to being recursive. Another suggestion be made was to restrict the grammars in such a way that the length of the derivation of any sentence's deep structure is less than x (some specified constant) times the length of its terminal string (his 'cut elimination' theorem). Peters and Ritchie (1973a) show that such a restriction generates the context sensitive languages.

Peters and Ritchie (1973a) take a somewhat different tack. Let f(x) be the maximal number of cyclic nodes[16] on any branch of the smallest base tree from which x can be derived (if x is in the language, otherwise $f(x) = 0$). If this function (the *cycling function*) is bounded by a recursive function, then the language is recursive. (For details, see Peters and Ritchie, 1973; Peters, 1973). One way to make f(x) recursive is via Peter's (1973) 'survivor property': the output of any cycle has more terminal nodes than any of its subparts on which the transformational cycle operated earlier in the derivation. Wasow (1978) suggests that this is inadequate for certain standard transformational derivations, and recommends replacing it by the 'subsistence property' (which is the survivor property with 'more' replaced by 'at least as many'), which continues to make f(x) recursive. Wasow gives various conditions on individual transformations which guarantee that the whole grammar obeys the subsistence property, e.g. that no specified deletion may effect more than one terminal node. In a similar vein, Lapointe (1977) gives a set of restrictions on deletion rules which guarantee the subsistence property.

One final example of the L-view is in Peters and Ritchie (1973b). In their (1971) they proved that every r.e. language could be generated from a context-free base language plus transformations. 'In proving [this] result we made heavy use of filtering, however' (1973b:180–181). A filter predicate is a symbol which may occur in the base language and be altered (moved, deleted, etc.) during the course of a transformational derivation. If such a symbol is still present at the end of the derivation, the sentence is 'filtered out' (said not to be generated by the grammar). In (1973b) they introduce 'local filtering': if the filter predicate is present at the end of any cycle, the sentence is not part of the language. The set of languages *C* produced by local-filtering transformational grammars from a context-free base is shown to contain all the context free languages, some but not all the context sensitive languages, and some non-recursive languages. 'Hence, *C* does not fit into well-known hierarchies of languages' (1973b:186). We have offered the above examples to emphasize our view that on the L-view, the goal of restricting the power of grammatical theories (and hence placing constraints on the permissible forms of grammars) is realistic, and to be preferred to the G-view. In the following sections we shall consider another kind of constraint, that of rule ordering, and investigate the generative capacity of grammars which invoke one or another rule ordering restriction. For example, one might insist that once a rule is applied in a given derivation it cannot be applied again. Or one might wish to require that there be a unique first rule that every output of the base language must go through first (if it can), and then this result is sent into a unique second rule, etc.

The question we wish to ask is: What effect does placing requirements (or restrictions) on the order of rule applicability have on the theory of grammar? Does it alter the class of languages produced? Does it alter the empirical predictions of the theory?

5. Rule orderings

In the earliest days of transformational grammar (Chomsky, 1955, 1957), the statement that the transformational rules of a grammar were ordered was a matter of definition. As late as 1962 Chomsky, in defending transformational grammars from the charge that they are 'merely taxonomic', says:

> In some sufficiently vague and general sense of the word "taxonomic", I have no doubt that this label can be applied to a transformational grammar. But I suggested "taxonomic" rather as a technical term to apply to a class of grammars based exclusively on segmentation and classification, without ordering of rules, and assigning only a single Phrase-Marker as full structural description on the syntactic level .... In this technical (and, I think, both useful and accurate) sense, a transformational grammar is not taxonomic. (p. 1002).

In more recent discussions, the status of the question of whether rules should be ordered by the linguist in describing the data (and if so, how ordered) has arisen from within transformational linguistics. Following standard linguistic usage, we shall call any theory of grammar which advocates (or allows) that rules of its grammars are to be ordered by the linguist, an 'extrinsically ordered theory', and any theory of grammar which advocates that rules of its grammars not have any order imposed on them by the linguist (although it may advocate certain principles which — in looking at some very general properties of the rules — dictate that some kinds of rules must apply before others in every derivation[17]) we shall call an 'intrinsically ordered theory'.[18]

We can distinguish three sorts of reasons offered in support of the view that rules of a grammar should (or should not) be ordered by the linguist in describing linguistic phenomena. All of these reasons, it seems to us, embody the G-view of constraints discussed in Section 4. The first reason we dub 'the aesthetics of rule ordering'. Arguing what seems to be the aesthetics of the matter, Koutsoudas *et al.* (1974) say:

> Predictions generated [within extrinsic ordered theories] by diachronic rule reordering were shown to follow simply from the more general, independently well-motivated principles of rule generalization and rule loss over time [in a non-extrinsic ordered theory]. (p. 26).

In a similar vein, Newton (1971) says 'if simplicity is the prime aim, then rules do not form ordered sets'. Pullum (1979:100) says 'an enormously strengthened theory is obtained if parochial [= extrinsic] ordering is completely forbidden', and that a theory with extrinsic ordering constraints 'contains a wholly unnecessary excrescence' (1979:27).

The same aesthetic taste can be found amongst advocates of extrinsic ordering, although they seem to think that extrinsic ordering is simpler. Dinnsen (1974) says:

> [Non-extrinsic orderings allow a wider class of grammars consistent with linguistic data.] Given two competing models of linguistic description where one model is more powerful than the other, in the sense that it has the effect of widening the class of grammars consistent with a given body of data, the burden of proof rests with the proponents of the more powerful model. (p. 33).

Soames (1974) says:

> If ... one's theory of grammar requires that every transformation be ordered with respect to every other transformation, then, since the two independently motivated transformations would have to be ordered with respect to each other anyway, selection of the rule ordering solution would be preferred because it gives us a chance to save ourselves the postulation of an extra grammatical device.

It would seem that we are well-advised not to base any of our theories on this aspect of the G-view. It is simply the case that no one knows what it is for a theory to be 'stronger' or 'simpler' when it comes to the issue of rule ordering. As we indicated in note 9, we think that responsible linguistics should not tolerate such *a priori* embracing or dismissal of linguistic theories.

The second sort of reason we find for advocating one or another (or no) ordering of the rules is avowedly psychological. Derwing (1973:212) calls ordering the rules 'a constraint-loosening strategem' in attempting to describe the actual psychological processes a speaker goes through. According to Derwing this is so because with ordering the linguist can write not only the (more 'complicated') intrinsically-ordered rules he advocates, but also could write 'simpler' rules in which the 'complications' are taken care of by ordering constraints. For similar remarks on how children 'really' only learn generalizations based on surface-data, and nowhere do they learn the 'very abstract representations which rule ordering constraints can give rise to' (see Skousen, 1975). Another trend along these same lines, apparently started by Perlmutter (1973) — as reported by Wasow (1975:375fn., 376fn.) — says that if two theories of grammar have a different number ways to order *n* rules, the one with the larger number 'invokes a wider class of grammars' and is therefore 'less preferable, a priori, all other things being equal'. Given a set of rules with *n* members, a theory of grammar which says that in any grammar the rules are not to be ordered will given rise to only one grammar, since there is but one way not to order *n* rules.[19] However, given a set of *n* rules, a theory of grammar which says that every grammar 'linearly orders' the rules will allow there to be *n*! distinct grammars. On the grounds of its 'invoking a narrower class of grammars', Perlmutter opts for no ordering restrictions. Wasow (1975) appears to follow Perlmutter in thinking that this is a good reason for rejecting certain theories of grammar, but nonetheless opts for 'linear ordering'. (Perhaps because he thinks the 'facts' dictate this, and so it is not true that 'all other things are equal'?) This argument reaches its culmination, at least in rhetorical force, in Pullum (1979:28–29), which we quote here in full. [The 'UDRA hypothesis' is the claim that rules are not to be ordered, except possibly by the above-mentioned 'universal principles' — see note 16].

> The number of distinct well orderings that can be imposed on a set of *n* elements is *n*!. Consider the question of how many grammars are defined as permissible by linguistic theory for a given set of *n* rules. Under the UDRA hypothesis the number is one. The interaction of the rules is universally determined, and given the *n* rules the nature of the corresponding grammar is fully determined. But under the hypothesis that parochial [= extrinsic] ordering constraints are admissible in grammars the number begins to rise rapidly, and attains its MAXIMUM under the "restrictive" strict linear ordering hypothesis: the number of grammars defined as permissible is *n*!. It is worth emphasizing the enormous extent to which this weakens linguistic theory. Burt (1971) provides a partial grammar for English which has 27 transformational rules. The UDRA hypothesis defines one grammar as the only one containing only these rules that is well-formed. (If the rules provided for any ungrammatical outputs, the claim would be that some of the rules must be wrongly formulated.) But the linear ordering hypothesis defines 27! (i.e. 27 × 26 × 25 ... × 1) grammars as well-formed for the same set of rules. This number is approximately ten thousand quadrillion.
>
> Such facts seem to me to justify the view that the linear ordering hypothesis should be regarded as an absurd one to adopt as a starting point. It should only be taken up under the most compelling evidence that no stronger hypothesis has a chance of success. It clearly does not provide a tough enough constraint on the formulation of rules; testing the claim "These rules R generate the language L" would involve, in the case of Burt's grammar, checking through ten thousand

> quadrillion orders of applications to see what approximations to English were obtained. It makes an absurdly weak claim about language acquisition: that even if Burt's rules were correct, and infants were born knowing the complete set of rules as well as all universal constraints defined by linguistic theory, they would still face the task of isolating the correct grammar for English from among a set of possible grammars of cardinality 200,000 million times greater than the number of seconds in the estimated age of the universe. What I am suggesting is that the same considerations that may be used to ridicule, quite correctly, the hypothesis that languages are learned by memorizing lists of sentences, also make the linear order hypothesis risible from an *a priori* standpoint.

As we said in Section 4, the rhetorical force of this kind of argument is diminished because it is hard to evaluate such psychological claims in the total absence of any theory of language acquisition. The rhetorical force is further dampened when one realizes that under the UDRA hypothesis, the child, in trying to evaluate whether he has the correct 27 rules will have to perform 27! possible derivations from each underlying structure. And finally, it should be pointed out that different theories will no doubt invoke different numbers of different rules in their grammars. It is far from obvious that a 'linear ordered theory of grammar' with 5 rules is 120 (= 5!) times worse than a 'UDRA theory of grammar' with 27 rules, if they both yield the same output.[20]

The third and final reason we find (both in the writings of the pro-extrinsic ordering linguists and in those of the anti-extrinsic ordering linguists) for preferring an ordering (or no ordering) of the rules is that there are actual cases of pairs of rules where, if one is ordered first, the grammar would produce ill-formed strings or else not produce all the possible strings (by not allowing the second rule to operate prior to the operation of the first rule). For a simple example, if we are given the string *ab* as an input to the two rules (1) $a \rightarrow d/\_\_b$, (2) $b \rightarrow d$, the order $\langle 1, 2 \rangle$ will yield the derivation $ab \rightarrow db \rightarrow dd$, while the order $\langle 2, 1 \rangle$ will yield the derivation $ab \rightarrow ad$. The language which is output from an input language containing only *ab* which linearly orders the two rules will therefore be either *dd* or *ad* (depending on which extrinsic order is chosen) but not both. Allowing these rules to be unordered will produce a language containing two strings, *dd* and *ad* from the input language *ab*. This is the strategy of many writers in linguistics on the topic (whether pro or con extrinsic ordering): they try to demonstrate that ordering (or not ordering) the rules of the linguistic theory won't give the correct output language (the natural language under discussion) and so the other option must be taken. Examples of this can be found in many writers, but see for example Ringen, Postal, Koutsoudas *et al.*, Dinnsen.

Now, many writers seem to think that such a demonstration is conclusive in determining whether the rules should be ordered and in which way, since presumably we know what the output is supposed to be (the well-formed sentences of the natural language). Unfortunately, the matter is not so simple as some of these writers seem to believe. For, not only is the ordering of the rules up for grabs, but so are the precise rules themselves and the input language. It would intuitively seem that one could make up for shortcomings of one of these by changing the others appropriately. For instance, in the preceding example of rules giving different languages, if we wanted the ⟨2, 1⟩ extrinsically ordered set to give us the same output as the intrinsically ordered rules do, we might make the input to the rules be the two strings *ab* and *dd*. Now, in actual linguistic cases it is seldom so simple to change the input or the rules, since a change in one place will have its effect in other areas of the grammar, but it is not implausible to believe that it can always be done by suitably changing the rules or input. And similarly, it is not implausible to believe that for any set of extrinsically ordered rules, there is a set of non-extrinsically ordered rules which yield the same language as output.[21] In any case, all of the writers canvassed above (except Chomsky in 1962) believe that the question of which theory of rule ordering should be adopted is 'empirical'. As Koutsoudas *et al.* (1974) put it:

> By showing that there is neither synchronic nor diachronic support for the hypothesis of extrinsic ordering, we have provided empirical support for the more restrictive hypothesis that all constraints on the relative application of phonological rules are determined by universal rather than language-specific principles of grammar. (p. 26).

Similar remarks can be found in writers from both sides of the dispute. We intend to investigate whether this is really an empirical dispute by investigating whether any of these theories differ in their strong generative capacity from any of the others. But first some preliminaries.

6. Some preliminary concepts

The question we wish to ask and answer is: what is the effect of ordering the rules of a grammar on its strong generative capacity? In spite of the enormous amount of writing on rule ordering recently, we can find only one writer who has even asked the questions; are any types of rule orderings stronger in either weak or strong generative capacity than any others? And, as we shall show below, there are some severe problems with that answer.[22]

One might wonder whether there is really any point in this exercise, since as we noted above (Section 4) transformational grammars can generate any r.e. set from any (context free) base; and this has been shown to hold regardless of whether the rules are ordered (as assumed by Peters and Ritchie[23]) or unordered (as assumed by Salomaa[24]). This remark is well-motivated, but the crucial thing to note about these proofs is that it is the principle of cyclic application which really does the work in these proofs. Our remarks here shall concern *intra*-cyclic ordering. Of course, even without cyclic application, theories of grammar with unlimited deletion rules and no requirements on the order of application of the rules can generate all r.e. sets. We therefore generally restrict our attention to those properties of theories of grammar which do not involve these factors. In fact, our discussion will often be concerned with grammars that have 'bounded' deletions (an upper limit on deletion for any one sentence). In the general case though, such factors will not enter into our discussion. The proofs we shall present hold for any non-cyclic sets of rule application. We shall be interested in proofs of the form: given an arbitrary (non-cyclic) grammar which obeys such-and-so rule ordering constraint, is there a (non-cyclic) grammar which obeys such-and-so *other* rule ordering constraint and which will produce the identical output if given the same input? That is, we shall be investigating the *relative* generative powers of theories of grammar that are otherwise identical but which invoke different rule orderings. This is, we believe, the most straightforward way to investigate the power of rule orderings *per se*, and not be investigating the power of the interaction of rule orderings with some other differing features of the grammars.

(One of these 'differing features' we shall not investigate is the effect of rules that are called (in linguistics) 'optional'. An optional rule would be one that needn't apply, but could. We could extend our treatment to cover optional rules (but won't here, see Pelletier and Fletcher, in preparation) by treating a grammar with an optional rule as though it were the union of two grammars: one with the rule (obligatorily) and one without the rule. Our study here is confined to 'obligatory' rules, for which we shall shortly give a precise definition.)

The 'preliminary notions' we are about to mention are given a rather 'formal' characterization. This is done so that those who wish to formally reconstruct our purposefully informal proofs of the 'theorems' in the next section will have the necessary tools at their disposal. Furthermore, such definitions give a precise, formal sense to our characterizations of the various types of rule orderings. For readers not interested in this formal sense, the informal characterizations given at the end of this section will afford sufficient content to understand the informal proofs of the next section. It should, however, be borne in mind that the rule ordering theories presented here are 'pure' theories which do not make recourse to any 'universal principles' on the application of the rules. Such 'universal principles' will be considered in Section 8.

We consider any structure in which there is a set of input, a set of output, a set of 'rules', and a statement of the 'legitimate application of the rules', to be a grammar. For 'set of input' we intend a set containing at least one symbol member (e.g. S for some PS grammars) or perhaps an infinite set of strings and their structure (e.g. the input the transformational component is the infinite set of structures generated by the base language). For 'set of output' we mean the strings of the language and their structure generated from the input by means of the 'rules' when 'applied properly'. We shall use I and L for 'input' and 'output language' respectively. Following Section 2, we say a *rule* is a certain kind of relation from strings and their structure to strings and their structure. We use P for sets of rules; and use p_i, p_j, ... for the members of P. We consider all sets of rules to be finite. A rule is said to be 'applied to a string X and immediately result in string Y' if X is in the domain of the rule and Y is the value of the production when applied to X.

We start by characterizing a very general theory of grammar. In this theory, a *derivation from I to L according to P* is a (possibly infinite) sequence of strings and structures such that (a) the first member of the sequence is a member of I, (b) if X is the n-th member of the sequence and Y is the (n + 1)st member, then there is some member of P such that it is applied to X and immediately results in Y, (c) if there is a last member of the sequence, there is no rule which it is in the domain of, (d) if there is a last member of the sequence, it is a member of L, and (e) if there is no last member of the sequence, the union of all partial initial segments of the sequence is a member of L.[25] We use ${}^{P}D^{1}$, ${}^{P}D^{2}$, ... to designate these sequences (derivations); we use ${}^{P}\mathscr{D}$ to designate the set of these sequences (the set of derivations from I to L according to P); and we use ${}^{P}\mathrm{d}^{i}_{1}$, ${}^{P}\mathrm{d}^{i}_{2}$, ... to designate, in order, the members of the sequence ${}^{P}D^{i}$ (that is, to designate the *stages* of the i-th sequence). The order of application of the rules in any particular ${}^{P}D^{i}$ is defined in the obvious way.

The class ${}^{P}\mathscr{D}$ of all derivations from I to L according to P is a well-defined set, as the remarks of the last paragraph indicate. Intuitively speaking, any rule is a candidate for application at any point in any derivation (whether or not it will actually apply depends on whether the other condition for a rule's applicability — its structural description — is met). This implies that there might be two distinct derivations ${}^{P}D^{i}$ and ${}^{P}D^{j}$ such that they are identical sequences up to their k-th member but where their (k + 1)st members are different. This will happen when the structural description of two distinct rules is simultaneously met by ${}^{P}\mathrm{d}^{i}_{k}$ $(= {}^{P}\mathrm{d}^{j}_{k})$. For every such case there will be two distinct derivations: one employing the one rule, one the other. And as we stated above, we restrict our attention to rules which would intuitively be called obligatory. In this general notion of grammar this restriction amounts to nothing more than our condition (c) of the last paragraph.

Given this characterization, we say that any rule is a candidate to immediately follow any rule. We shall call this theory of grammar which imposes no ordering restrictions a *random ordering* of the rules. (This is what Koutsoudas, 1976b, calls 'The Random Sequential Hypothesis'). In linguistics, which rules are candidates for application at a given stage of a derivation are often restricted in some way. For example, given a set *P* of rules, certain pairs may be selected to form a set $\mathscr{F}$(P), the set of ordered pairs of rules where one rule (the second) may immediately follow the other. Such sets of pairs of rules can be used to formulate a class of 'first rules' — those rules which are candidates for application to the first stage of any derivation. Given that $p_i \in P$ we say:

p_i is a first rule of $\mathscr{F}$(P) iff:
either (1) p_i is the only rule of *P* or (2) if there is any rule in *P* that is allowed to precede p_i, p_i can also precede it. I.e. for any p_j, if $\langle p_j, p_i\rangle \in \mathscr{F}$(P) then $\langle p_i, p_j\rangle \in \mathscr{F}$(P).

There are many kinds of $\mathscr{F}$(P) sets that *could* be constructed from a set *P* of rules; we shall restrict our attention to two — (a) those sets which impose an asymmetric relation on *P* and (b) those which impose an antisymmetric relation on *P*.[26] We shall denote then by $\mathscr{F}^1$(P) and $\mathscr{F}^2$(P) respectively; when no superscript is used, we mean the claim to be true of each. Intuitively, proponents of theories which invoke such restrictions want to say that all legitimate derivations are such that at any stage, the only way to get to the next stage is to use a 'next rule' after the one just previously used, or, if that state doesn't meet the structural description of that 'next rule', then use the very 'next' one, ... etc. We formalize this intuition with help of the 'immediately applicability ancestral with respect to $^{P}d^{n}_{m}$', symbolized $\mathscr{A}(^{P}d^{n}_{m})$:

1. if p_i is a first rule of $\mathscr{F}$(P), then $p_i \in \mathscr{A}(^{P}d^{n}_{1})$
2. if $^{P}d^{n}_{m-1}$ immediately results in $^{P}d^{n}_{m}$ by application of p_j and $\langle p_j, p_i\rangle \in \mathscr{F}$(P), then $p_i \in \mathscr{A}(^{P}d^{n}_{m})$
3. if $p_j \in \mathscr{A}(^{P}d^{n}_{m})$ and $\langle p_j, p_i\rangle \in \mathscr{F}$(P) and $^{P}d^{n}_{m}$ does not meet the structural description of p_j, then $p_i \in \mathscr{A}(^{P}d^{n}_{m})$.

All derivations from *I* to *L* according to *P* are said to be legitimatized by $\mathscr{F}$ and $\mathscr{A}$ if and only if (1) the first member is a member of *I*, (2) if *X* is the n-th member and *Y* is the (n + 1)st member, then there is some member of

$\mathscr{A}(X)$ which is applied to *X* and immediately results in *Y*, (3) if $p_i \in \mathscr{A}(X)$ and X meets the structural description of p, then no ${}^{P}D^{k}$ has X as its final stage.

While there are various kinds of procedures one might use to set up an ordering of the rules, the discussion of the preceding paragraph naturally lends itself to orderings of the rules which have these two features: (1) the order of application $\langle p_i, p_j \rangle$ is legitimate in one derivation if and only if it is legitimate in all derivations (a "transderivational constraint" on rule orderings), (2) given that p_i has applied in a derivation yielding stage ${}^{P}d^{n}_{m}$, one knows what the next rule(s) permitted to apply is (are) — viz., any rule which is an element of $\mathscr{A}({}^{P}d^{n}_{m})$ — a 'local', 'derivational constraint'.

In addition to the random ordering discussed above (which vacuously fits these conditions), there are two natural types of rule orderings which fit this picture: *total orderings* — where there is a unique first rule, a unique second rule, etc.; and *partial orderings* — where there is a unique first rule, but thereafter there are two rules which are candidates for application at that stage, the one which just applied and the (unique) next different one. (So partial orderings are like total orderings except that at any point in a derivation one is allowed to apply again the same rule which just got applied.) The formal difference between the two is whether $\mathscr{F}^1(P)$ or $\mathscr{F}^2(P)$ is used in defining $\mathscr{A}({}^{P}d^{n}_{m})$. Total orderings are by far the most common presentation in the literature on transformational syntax and phonology, especially in introductory textbook exposition. Partial orderings are not so well represented in the literature, but it is argued for by Anderson (1969:85–87), Kenstowicz and Kisseberth (1973), Johnson (1972) and Palacas (1971) as a way of showing shortcomings in an opposing theory of application (which we shall discuss in Section 8) put forward by Chomsky and Halle (1968:344). In this literature this kind of rule is called 'iterative'.

We might relax condition (2) of two paragraphs ago and allow ordering restrictions to be stated as a non-local, but still 'derivational', constraint. That is, we may be interested in eliminating or retaining derivations on the basis of restrictions which are not stated merely in terms of rules that relate *adjacent* stages of a derivation, but rather impose a restriction which relates non-adjacent stages (a 'global derivational constraint'). For example, we might want to impose the restriction of asymmetry: if some derivation allows the rules to apply in the order $\langle p_i, p_j \rangle$, then no derivation allows the order $\langle p_j, p_i \rangle$, although p_j might be applied without applying p_i. We call this a *semi ordering* of the rules. Or we might want to impose the restriction that if some derivation allows the rules to be applied $\langle p_i, p_j \rangle$, then no derivation allows the rules to be applied $\langle p_j, p_i \rangle$ unless $i = j$, although p_j can be applied without applying p_i. We call this a *semi* *partial ordering* of the rules. Anderson's (1969:12) definition of 'local ordering' theories is a way of stating our semi partial or semi orderings (depending on whether one allows rules to apply 'iteratively' to their own output or not, respectively).

There are various equivalent ways of formally stating the class of derivations legitimatized according to this intuitive characterization. One simple way, making use of the concepts already developed, is to slightly alter the 'immediately applicability ancestral' so that derivations need not start with a "first rule" of $\mathscr{F}$(P), nor need it always use a "next" rule of $\mathscr{F}$(P). This new ancestral we indicate as $\mathscr{B}({}^{P}d^{n}_{m})$.

1. if $p_i \in P$, then $p_i \in \mathscr{B}({}^{P}d^{n}_{1})$
2. if ${}^{P}d^{n}_{m-1}$ immediately results in ${}^{P}d^{n}_{m}$ by application of p_j and $\langle p_j, p_i\rangle \in \mathscr{F}(P)$, then $p_i \in \mathscr{B}({}^{P}d^{n}_{m})$
3. if $p_j \in \mathscr{B}({}^{P}d^{n}_{m})$ and $\langle p_j, p_k\rangle$, $\langle p_k, p_l\rangle \ldots\langle p_x, p_i\rangle$ are all elements of $\mathscr{F}$(P), then $p_i \in \mathscr{B}({}^{P}d^{n}_{m})$.

The characterization of all derivations from *I* to *L* according to *P* which are legitimatized by $\mathscr{F}$ and $\mathscr{B}$ is just as before except that $\mathscr{B}({}^{P}d^{n}_{m})$ replaces $\mathscr{A}({}^{P}d^{n}_{m})$. Again, the difference between semi orderings and semi partial orderings is whether $\mathscr{F}^1$(P) or $\mathscr{F}^2$(P) is used in defining $\mathscr{B}({}^{P}d^{n}_{m})$.

We also want to consider a relaxation of condition (1), and not force ordering restrictions to be transderivational constraints. For example, we may want to require that each derivation apply the rules asymmetrically, but permit different derivations to do it in a different order, so long as each one applies them asymmetrically. We call such ordering restrictions *unorderings* of the rules. An unordering of the rules is what Koutsoudas (1976b) calls 'The Arbitrary Ordering Hypothesis' and what is implied by Ringen's (1976) PRINCIPLE VI (unrevised). Or we may want to impose the condition of antisymmetry on individual derivations. We call this a *quasi ordering* of the rules. This sort of ordering is implied by Ringen's (1976) PRINCIPLE VI (revised). These two restrictions are not easily stated in terms of the concepts already developed, since $\mathscr{F}(P)$, and hence the immediate applicability ancestral, are stated as giving a condition which holds for all derivations. We shall here forego stating the formal apparatus necessary to define unorderings and quasi orderings — their intuitive content is clear enough. (See the end of this section for a statement of them.)

A final kind of condition on application of a set of rules we wish to consider is what has been called *simultaneous application* of the rules. (See Chomsky and Halle 1968:19 footnote 5. Koutsoudas, 1976b, calls this 'The Direct Mapping Hypothesis'. It asserts that there are no intermediate representations between an underlying form and its corresponding surface form in any derivation.) This conception of rule application is very ill-understood, in spite of the fact that it is commonly invoked, especially in the phonological literature. One problem is that two rules which are supposed to be applied simultaneously may both have their structural description met by *X* and each rule require a change in *X* which cannot simultaneously be done (i.e. require contradictory changes).[27] We shall charitably assume, for the rest of this paper, that either the sets of rules are chosen so as to never have this happen or for such cases to have an output defined for them. (We shall shortly give a criterion for non-contradictoriness.) The idea behind simultaneous application of the rules is to have all the rules 'apply at once', so as not to have structure created or destroyed (which would otherwise affect how the rules operate). Consider, for example, this set of rules:

1. c → b /__a
2. a → b /c__
3. b → d /b__

With the input *cab* we should arrive at *bbb* by simultaneous application. But no order of 1–3 will give it. (For example, the order ⟨1, 2, 3⟩ gives *bab*). The intuitive content of 'simultaneous application' is clear, at least if the rules do not require contradictory changes. One way of ensuring non-contradictoriness of simultaneously applied rules would be to require it always to be possible to apply the rules to any input string by using any order of the rules and constructing a piece-by-piece 'copy' of the output, and that the final output of such a derivation will be the same for any order of the rules. For our simple example above, such a process would look like this (choosing the order ⟨1, 2, 3⟩):

input temporary copy *final*

$$\left.\begin{array}{l} \text{rule 1: cab} \rightarrow \text{b - -} \\ \text{rule 2: cab} \rightarrow \text{- b -} \\ \text{rule 3: cab} \rightarrow \text{- - - (doesn't apply)} \end{array}\right\} \text{bbb}$$

The idea is that every rule operates on the same input, namely the relevant member of *I*, but we build up the output piece-by-piece as a temporary copy. We keep adding to the temporary copy until we have tested all the rules. If every order of application of the rules yields the same final output, then the simultaneously applied rules are not contradictory. (Reflection will show that different orders yield different results in the examples of contradictory rules given in the previous footnote.)

We close this section with an informal description of the eight types of rule ordering constraints which might be imposed on otherwise identical theories of grammar, thus yielding us eight distinct theories of grammar.

We use the terminology that a theory invokes a rule ordering if it would deny legitimacy to certain derivations because of the order in which the rules of that derivation applied. (We vacuously include random ordering as one of these theories; we also include simultaneous application of rules.)

Total Orderings: there is a unique first rule, a unique second, ... a unique last rule. (See standard texts for examples.)

Partial Orderings: there is a unique first rule, but thereafter at every stage of a derivation there are two rules which are candidates for application: the rule which was just applied and the (unique) next different rule. (See Johnson's 1972 definition of 'iterative rule'.)

Semi Orderings: the rules are given a total ordering, but different derivations may start at different places in the ordering (and choose any 'later' rule as the next rule applied). (Anderson's 1969 'local ordering' without 'iterative rules'.)

Semi Partial Orderings: the rules are given a partial ordering, but different derivations may start at different places in the ordering (and choose any 'later' rule as the next rule to be applied). (Anderson's 1969 'local ordering' with 'iterative rules'.)

Unorderings: any derivation can apply the rules in any order, subject only to the constraint that once a rule has been applied in a derivation, it is no longer eligible for application at a later stage. (Koutsoudas's 1976b 'arbitrary order hypothesis' and Ringen's 1976 PRINCIPLE VI, unrevised.)

Quasi Orderings: any derivation can apply the rules in any order, subject only to the constraint that once a rule has been applied in a derivation, the only other time it may be applied is to its own output. (Ringen's 1976 PRINCIPLE VI, revised.)

Random Ordering: there is no order imposed on the rules; any derivation can apply the rules in any order. (Koutsoudas's 1976b 'random sequential hypothesis'.)

Simultaneous Application: the entire set of rules is applied to an input 'all at once'; this prevents some of the rules from creating or destroying part of the input in such a way as to affect the applicability of other rules. (Koutsoudas's 1976b 'direct mapping hypothesis'.) (The text above discusses the notion of 'contradictory rules' in the context of simultaneous application, and gives a condition which guarantees the non-contradictoriness [for simultaneous application] of a set of rules).

In passing, we should note one further feature of our definitions. If, for example, P is randomly ordered, no member of ${}^{P}\mathscr{D}$ is illegitimate because of an ordering restriction. However, this is *not* to say that there will be members of ${}^{P}\mathscr{D}$ which have, say, p_i actually apply before p_j in one derivation and p_j actually apply before p_i in another. For it may be the case that, given *I* and *P*, every derivation in $^P\mathscr{D}$ happens to apply the rules in the same order (because of the structural descriptions of the rules and the nature of the members of *I*). That is, the rules might be *intrinsically ordered* to behave as if they were totally ordered. Or, they might be intrinsically ordered to behave as if they were semi-partially ordered. We emphasize this point because we want to be certain that it is clear that this is still a random ordering of the rules: no restrictions have been placed on the appropriate order of application of rules in any derivation. No possible member of $^P\mathscr{D}$ will be ruled out by an ordering restriction.

Finally, we remind the reader that we are concerned only with intracyclic strength of grammars — that is, we have no cyclic application of rules; and furthermore, we are concerned only with 'obligatory rules'. (See above for discussion of 'optional rules'.)[28]

The strong generative capacity of the eight rule ordering theories

We shall now state a series of theorems which are summarized in Figure 1. These theorems tell us the relative strong generative capacity of the 'pure' theories of rule ordering defined in Section 6. ('Pure' in the sense of involving no interaction between the rule ordering principles and any other principles.) We shall postpone the proofs of most of the theorems (or rather: the instructions for constructing proofs) to the Appendix. These proofs are all 'constructive' in the sense of giving a method for actually exhibiting a grammar having certain (rule ordering) properties. While we postpone proofs of most, we illustrate the general procedure in the text by showing how Theorem 1 is to be proved. In the following Section 8, we consider some 'impure theories' which have been proposed in the linguistic literature.

Our strategy in these proofs is this: We start with a theory of grammar which says that all grammars X-order the rules. We shall try to construct a grammar which Y-orders the rules yet gives all the identical empirical predictions as any X-ordered grammar (because it gives exactly the same derivations). This will prove that Y-orderings are at least as strong in strong generative power as X-ordering theories of grammar. We then will show in what cases the converse cannot be established. The intuitive fact which allows the proofs to go through is that every derivation of a weaker theory is the initial segment of some derivation of the stronger theory. We therefore need only a method of 'stopping derivations appropriately', and a method of preventing other derivations from getting started. We illustrate Theorem 1 in some detail and make comments about the proof to show that the grammar exhibited as constructed from the totally ordered grammar actually obeys standard linguistic conventions on the form of rules, etc. The strategy for proving Theorems 2–7 is essentially that of Theorem 1 with suitable alterations in the technique of 'keeping track of the rules' to account for the different kinds of rule orderings.

Theorem 1: Partial orderings, semi orderings, semi-partial orderings, quasi orderings, unorderings, and random orderings are all at least as strong as total orderings.

We start by showing that, for every L generated from an I according to some P in accordance with $\mathscr{F}^1(\mathrm{P})$ and $\mathscr{A}$, there is a random ordering of some set of rules such that L and only L can be generated from I according to that set of rules. Suppose we have a totally ordered set of n rules. Then there is a unique first rule which, for illustration, let us suppose is the following phonological rule.[29]

1. $$\begin{bmatrix}+\text{high}\\-\text{round}\end{bmatrix} \rightarrow \begin{bmatrix}+\text{round}\\+\text{front}\end{bmatrix} \quad /[+\text{back}]\_\_\_[+\text{high}]$$

Since this is the first rule, every member of the input must first be checked to see if this rule applies, before any other rule can be considered. In constructing a randomly ordered set of rules which will have the same effect as this first rule, we give two rules which involve one new (terminal) symbol. For simplicity we call the symbol [+1]. One of the new rules will tell us the effect of the rule's applying non-vacuously:

1a. $$\begin{bmatrix}+\text{high}\\-\text{round}\end{bmatrix} \rightarrow \begin{bmatrix}+\text{round}\\+\text{front}\\+1\end{bmatrix} \quad /[+\text{back}]\_\_\_[+\text{high}]$$

The other rule gives us the effect of the rule's not 'really' applying (when either the structural description or the context requirements are not met).[30]

1b. $$\mathrm{A} \rightarrow \begin{bmatrix}\mathrm{A}\\+1\end{bmatrix} \quad /\alpha\_\_\_\beta \quad \textit{cond}\text{: either } \alpha \neq [+\text{back}] \text{ or } \beta \neq [+\text{high}] \text{ or } \mathrm{A} \neq \begin{bmatrix}+\text{high}\\-\text{round}\end{bmatrix}$$

To construct the randomly ordered analogue of the second rule of the total ordering we again construct two rules: one for 'real' and one for 'non-real' application of the rule to the string going through them. The structural description of these rules will have to specify that the string has gone through one of the previously-constructed two rules. We insure this by adding to the structural description of the two new rules the symbol [+1]; the structural change of these rules will change [+1] to [+2]. In other respects they resemble what 1a and 1b did to rule 1. For example, suppose the second rule of the totally ordered grammar were:

2. $[+\text{back}] \rightarrow [-\text{back}]$ $/[+\text{front}]$____

Our pair of rules of the randomly ordered grammar would be:

2a. $\begin{bmatrix}+\text{back}\\+1\end{bmatrix} \rightarrow \begin{bmatrix}-\text{back}\\+2\end{bmatrix}$ $/[+\text{front}]$____

2b. $\begin{bmatrix}\text{A}\\+1\end{bmatrix} \rightarrow \begin{bmatrix}\text{A}\\+2\end{bmatrix}$ $/\alpha$____ *cond*: either $\text{A} \neq [+\text{back}]$
or $\alpha \neq [+\text{front}]$

Every rule of the totally ordered set is replaced by two such rules as these. We claim that this randomly ordered set of rules will, given the same *I*, generate the same *L* as the totally ordered set. The output of the rules corresponding to the n-th (last) ordered rule will have (among other things) the feature [+n]. Note that we cannot simply delete this symbol, since then the output would be eligible to ‘start the rule-sequence over again’ because the structural descriptions of the rules corresponding to the first totally ordered rule will be met. Levine (1976:119) is wrong at this point. His method amounts to dropping the ‘method of keeping track’ at the end of the randomly ordered derivation; therefore, there will be cases where the end product will be eligible to ‘start the rule sequence all over again’ or to start somewhere in the ‘middle’ of the rules. So his conversion procedure will not yield the identical language, and is therefore incorrect. There are three ways to avoid this difficulty (all of which amount to the same thing) to convince one that the output here is the same output as the totally ordered rules gave: (1) if one's interest is purely formal this will suffice: define the language produced by the rules to be those strings/structures without the [+n] markers; (2) if one's interest is in phonology, this will sound more familiar: introduce a convention to the effect that [+n] is not phonetically realized; (3) if one is a syntactician, this will be familiar: replace [+n] by the null element and distinguish sentences where some non-terminal node dominates *null* from sentences without that non-terminal node. (This course is not so familiar in phonology, but see Dinnsen, 1972, 1974 for an example of it.)

This proves that random orderings are at least as strong as total orderings in strong generative capacity. Now note that this newly-constructed set of rules could have been required to be quasi ordered, unordered, semi-partially ordered, semi ordered or partially ordered, and still their intrinsic ordering will be the same as the original total ordering. Hence Theorem 1 is proved.

Let us make a few remarks on the conversion procedure. First we note that these new rules are all legitimate rules of linguistics if the totally ordered ones we started from were. What we have done is add a series of [+i] markers; but the [+i] markers are here employed only for mnemonic convenience. What we have done is show that there is *some* randomly ordered set of rules for the language. In fact there are an infinite number. In particular, some of these sets of rules will have ordinary-looking symbols in place of our [+i] markers. These ordinary-looking symbols may even be used elsewhere in the language, if it so happens that their effect in the rules is exactly what the [+i] markers' effect is — to make sure that segments go through the rules 'in the right order'. In fact, since in most of the totally ordered grammars which have been given in the literature, many rules could be placed in various positions in the ordering, and since most other rules are already intrinsically ordered with respect to one another (for illustration of these facts, see Pullum, 1979, for syntax and Derwing, 1973, for phonology), the ordinary-looking markers mentioned in the last sentence don't even have to do a perfect job. Just 'keep track' where it's important. Examples of this happening are provided by many syntactic transformations. Consider PASS in the formulation given by Bach (1974): the application of PASS is the only way to get *by* in precisely the position it occupies in the structural change of that rule — no other transformation or base rule will do it. In this transformation, *by* plays the role of one of the [+i] markers. Consider also his formulation of AFFIX SHIFT: this rule introduces word boundry symbols, ⊕, into the output. Since this is the only rule which introduces ⊕, any string with ⊕ in it must have gone through AFFIX SHIFT, and therefore any rule with ⊕ in its structural description is intrinsically ordered after AFFIX SHIFT. Postal's [+doom] markers provide another example, as do certain formulations of trace theory (see, e.g. Lightfoot, 1977; Fiengo, 1977; Chomsky and Lesnik, 1977). Similar remarks can be made about many transformations — so it's certainly not 'violating the form of transformational rules' to talk about [+i] markers.

Now consider the 'b' rules of the random ordering — the ones corresponding to non-application of the totally ordered rules. The method we have given is *general* in the sense that it will apply to any totally-ordered-randomly-ordered conversion. In a particular grammar there are various ways the statement of the 'b' rules, expecially the 'condition' part, might be simplified and retain the same effect. (See also note 29 for other remarks on the statement of conditions.) For example, the non-identities there might be replaced by an identity to some other feature or element. The possibility of doing this depends upon whether the grammar of the language under consideration has a way of specifying 'what it is to be non-identical with [+high, – round]' (for our rule 1b). If there is such a feature or element in the grammar, say for example γ, then the condition on our rule 1b could be put in terms of '$A = \gamma$'. This would make the condition look more like actually proposed conditions on transformations. The point here is that, given a condition like '$A = \gamma$', one doesn't know whether it 'performs merely a keeping track function' or 'really does some work in the way any ordinary rule does'. More generally, we would say that there is no way for an outside linguist to tell, given only the 'form of the rules', whether a certain grammar was constructed by a linguist whose aim was to have a 'genuinely' randomly ordered set of rules or was constructed by a linguist who intended to have a totally ordered set of rules but had someone alter them along the lines here stated.

Finally we note that this conversion preserves *strong* generative capacity — any empirical claim (paraphrase, synonymy, surface ambiguity, deep ambiguity, etc.) made by totally ordered grammars is also made by this randomly (or quasi, or un-, or semi, ...) ordered set of rules. This is because the derivations in the two theories are almost identical. We note that at any stage in a derivation in the randomly ordered set of rules, exactly one rule will apply. Every stage of the totally ordered derivation has an identical stage (but for the presence of [+i] markers) in the randomly ordered derivation. There are perhaps a few extra stages in the randomly ordered grammar, ones which 'mark time' until the next rule which applies comes up. But this is not to alter any empirical claim that the totally ordered grammar made. Indeed, it's not to make any empirical claim at all; it is merely an artifact of the grammar. In every respect the two grammars generate the same surface structure from the same members of *I*, and they are related to intermediate stages in the identical way. So any claim which can be made about surface strings, surface structure, or derivational history in the one can be made in the other. Thus for any totally ordered grammar there is a randomly ordered (quasi ordered, semi ordered, etc.) grammar which will make precisely the same empirical claims about the language.

We trust therefore, that we have convinced the reader that our use of markers like [+i] is not only permitted by the conditions placed upon linguistic rules, but that they are in fact indistinguishable from actual linguistic practice. The comments of Ringen, Koutsoudas *et al.*, and Pullum (1979b:185, footnote) to the effect that this 'indexing of rules' is somehow a trick which can always be spotted (and that it is 'really' the same as totally ordering the rules), is thus shown to be false.

Theorem 2: Random orderings, quasi orderings, and semi partial orderings are all at least as strong as partial orderings.

Theorem 3: Random orderings, quasi orderings, unorderings, and semi partial orderings are all at least as strong as semi orderings.

Theorem 4: Random orderings and quasi orderings are each at least as strong as unorderings.

Theorem 5: Random orderings and quasi orderings are each at least as strong as semi partial orderings.

Theorem 6: Random orderings are at least as strong as quasi orderings.

Theorem 7: Total orderings are at least as strong as simultaneous application.

We now turn our attention to the question of whether the 'at least as strong' in Theorems 1–7 can be replaced by 'stronger'. We sketch a proof of Theorem 8 and merely state the others; their proofs are in the Appendix.

Theorem 8: Random orderings, quasi orderings, unorderings, semi orderings, semi-partial orderings, and partial orderings are all stronger than total orderings.

All total orderings have the property that there are no more members of the output language than there are in *I*; and this is independent of which rules are chosen. Except for partial orderings, each of the other orderings mentioned have more than one 'first rule', and so for each of them it is possible to give a set of rules which will generate from a given input member, more than one output. Partial orderings have a unique 'first rule', but thereafter there are two rules which are candidates for application. Thus two distinct derivations can start from the same member of *I*. But this can never be done in a total ordering (with obligatory rules). Hence, all of these orderings can produce languages from some *I* that total orderings cannot, and can make differing claims about the relation between members of *L* and *I* from what total orderings can claim. Together with Theorem 1, this proves Theorem 8.

Levine (1976:120f) says that random and total orderings are equivalent because 'every derivation follows some order or other'. He gives this example (p. 121):

> ... let us look at an unordered [= our random] system consisting of three transformations, A, B, C. The only possible orders of application for these transformations are
> 1\. ABC 2. ACB 3. BCA 4. BAC 5. CAB 6. CBA
> These choices can be accounted for by having an ordered system with the following indexed list of transformations [= method of keeping track of the rules, analogues to the present [+i] markers]:
> $\mathrm{T}_1 = \mathrm{A}\ \mathrm{T}_2 = \mathrm{B}\ \mathrm{T}_3 = \mathrm{C}\ \mathrm{T}_4 = \mathrm{A}'(= \mathrm{A})\ \mathrm{T}_5 = \mathrm{B}'(= \mathrm{B})\ \mathrm{T}_6 = \mathrm{C}'(= \mathrm{C})\ \mathrm{T}_7 = \mathrm{A}''$ $(= \mathrm{A})$.
> Then we can rewrite 1–6 above as follows:
> 1\. $\mathrm{T}_1\mathrm{T}_2\mathrm{T}_3$ 2. $\mathrm{T}_1\mathrm{T}_3\mathrm{T}_5$ 3. $\mathrm{T}_2\mathrm{T}_3\mathrm{T}_4$ 4. $\mathrm{T}_2\mathrm{T}_4\mathrm{T}_5$ 5. $\mathrm{T}_3\mathrm{T}_4\mathrm{T}_5$ 6. $\mathrm{T}_3\mathrm{T}_5\mathrm{T}_7$ etc.

However, this is incorrect. No order can be given to these seven T-rules. For, suppose every member of the input language happens to meet the structural description of rule A ($=T_1$). It is still the case, in the randomly ordered grammar, that (say) option 3 is open — after all, not all derivations have to start with rule A, even if all members of *I* meet its structural description. However, in Levine's totally ordered grammar every member of the input will meet the conditions for T_1 and therefore will be changed. Such a change might prevent anything from going through rule B ($=T_2$) at all, much less going through it first. So this portion of Levine's 'equivalence proof' is totally incorrect; randomly ordered grammars *are* stronger than totally ordered ones.

Theorem 9: Semi partial orderings are stronger than semi orderings.
Theorem 10: Quasi orderings are stronger than unorderings.
Theorem 11: Random orderings are stronger than quasi orderings.
Theorem 12: Semi partial orderings are stronger than partial orderings.
Theorem 13: Unorderings are stronger than semi orderings.
Theorem 14: Quasi orderings are stronger than semi partial orderings.

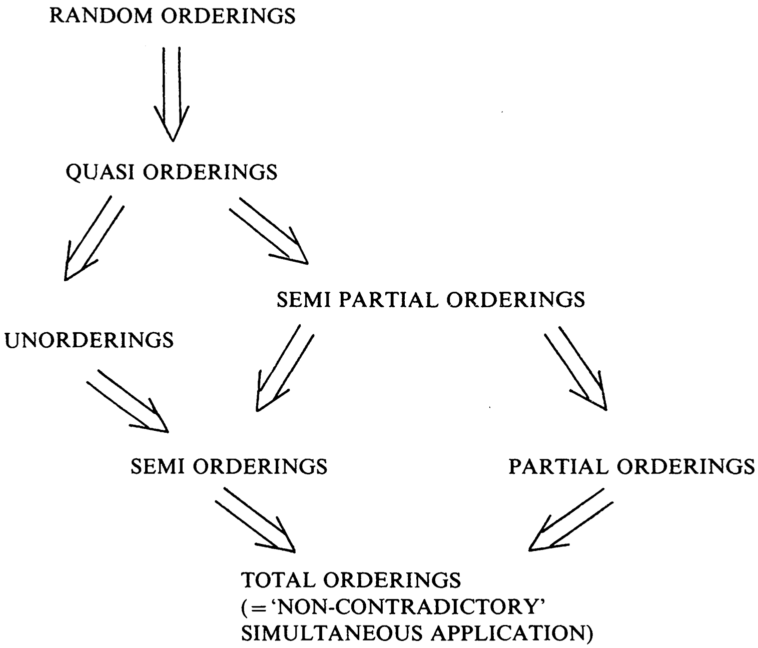

Figure 1. *Relative Strong Generative Capacities of Various Theories of Rule Orderings. $X \rightarrow Y$ means that theory X is stronger than theory Y. ($\rightarrow$ is transitive)*

Theorem 15: Total ordering is equivalent to simultaneous application.
Theorem 16: Neither unorderings nor semi orderings are comparable to partial orderings.
Theorem 17: Unorderings and semi partial orderings are not comparable.

Theorems 1–17 give us the relationships illustrated in Figure 1. We take these theorems to show that Chomsky and Halle are wrong when they say (1968:18):

> It is always possible to order the rules in a sequence and to adhere strictly to this ordering in constructing derivations without any loss of generality as compared to an unordered set of rules or a set ordered on a different principle.

Or, wrong at least if they are talking about the strong generative capacity of theories of rule ordering.

8. 'Universal principles' on rule application

Many linguists would be unhappy with our characterization of a derivation. In particular, they would be unhappy with the possibility of a derivation not having a last member (and our decree that in this case the member of the output language be the union of all initial sequences of the derivation), and they might try to eliminate this possibility in various ways. The most popular way would doubtless be to try to restrict the 'form of the rules' in some way or other. (After all, they might point out, one way a randomly ordered derivation might not have a final member would be to have rules like $a \rightarrow ab$.) However, merely disallowing rules like this would not solve the difficulty, for we can also exhibit pairs like $a \rightarrow bc; b \rightarrow ad$. And the 'circle' exhibited here could be expanded to whatever size the set of rules is. So we think that merely arguing about 'the form of rules' is inadequate. And in any case, it would not touch the proofs of the previous section, since restricting the 'forms of rules' would no longer have us just investigating the strong generative capacity of various types of rule orderings, but rather what effect *this* restriction would have on rule ordering strength. We can think of two ways of invoking our rule orderings in such a way that it will extricate advocates of (what they call) 'random ordering' from this difficulty. We don't think many such advocates have ever done this, but we attribute this to their lack of thought on the possibility of having derivations without last members; for, their theories of grammar do not rule this out.[31] Levine (1976:121f) discusses some possible constraint to the effect that the output of any sequence of transformations cannot be identical to any part of its input — a constraint which seems to him to solve the problem. However, he seems not to notice either that (a) this shows that his earlier equivalence proof must be wrong (since this restriction on random ordering is necessary), nor (b) that the addition of such a constraint really is a *constraint*, which powerfully restricts the class of grammars permitted by 'random ordering'. In any case, it is still inadequate unless some caveat about variables is added, for in both of our examples of troublesome rules, $a \rightarrow ab$ and the pair $a \rightarrow bc$, $b \rightarrow ad$, the rules do not have outputs identical to the input nor to part of the input. Various of the obvious things to say about stating rules with variables seem to rule out standard formulations of many transformations like AFFIX SHIFT, and most phonological rules. Such a condition would therefore seem to be not acceptable to advocates of 'random ordering'.

One way to avoid this difficulty would be to claim that for every rule there is a finite number of times it is to be considered available for application in any derivation. Other than this restriction, anything legitimate in a derivation of a random ordering is legitimate here also. Note that the effect of this is to invoke an unordering of the rules except that each rule is allowed to apply (not just once but) some finite number of times.[32] We shall want to say that this is still an unordering of the rules, and we subsume it to our original definition by adding, for each rule, the appropriate number of copies. (And now each copy is allowed to apply only once, as in an unordering.) We imagine also that we have some marker to distinguish different occurrences of what is otherwise the 'same' rule. We call adding 'copies' of rules already in P, the forming of a $^*P$ set of rules. And we shall call this kind of unordering a $^*P$-unordering of the rules P. Proposal A is one way of understanding what linguists who advocate (what they call) 'random ordering' of P must have in mind to avoid the above-noted difficulties. This seems to be the position advocated in Koutsoudas *et al.* (1974) and Ringen (1972).

Proposal A: Theories of grammar should be (equivalent to) *P-unorderings of the rules.

A second way these writers who advocate 'random ordering' might restrict their rules in order to avoid having derivations with no last member would be to invoke quasi orderings and decree that no rule is allowed to create (any part of) its own structure. So, we eliminate rules of the form $a \rightarrow ab$, etc., and impose a quasi ordering. This seems to be advocated by Ringen (1976) in her revised PRINCIPLE VI (and surrounding discussion). We call sets of rules that have the form $+P$ sets, and take these theorists to be advocating

Proposal B: Theories of grammar should be (equivalent to) $+P$ quasi orderings of the rules.

We now investigate two linguistic proposals for 'universal principles of rule application' which have been made by advocates of 'random ordering'; we show that each of them, when added to a $*P$-unordering is equivalent in strong generative capacity to total ordering. And then we show that each of them, when added to a $+P$ quasi ordering is intermediate between total and partial orderings. (The proofs of Theorems 18–23 are in the Appendix.)

The first proposal, Universal Principle I, appears in Ringen (1972) who claims that every 'genuine' set of (what she calls) 'randomly ordered rules' will never be such that some stage of a derivation can simultaneously meet the structural description of two distinct obligatory rules. That is to say, she thinks it is somehow contained in the concept of 'obligatory rule' that such a state of affairs never be permitted. (See also her 1976 and the discussion in Hastings, 1976.) We call Ringen's condition:

Universal Principle I: No stage of a derivation can simultaneously meet the structural descriptions of more than one obligatory rule.

Another proposal, as a condition on rule application in a 'randomly ordered theory', has been made by Koutsoudas *et al.* and by Derwing (1973).[33] It is the position that a stage of a derivation *can* simultaneously meet the structural description of more than one obligatory rule, but that when it does, all these rules are simultaneously applied to reach the next stage. We call this KSN/Derwing condition on rule application:

Universal Principle II: If a stage of a derivation simultaneously meets the structural descriptions of more than one rule, the rules are to be applied simultaneously.

Theorem 18: Theories of grammar obeying Proposal A and Universal Principle I are equivalent to total orderings.

Theorem 19: Theories of grammar obeying Proposal A and Universal Principle II are equivalent to total orderings.

Theorem 20: Theories of grammar obeying Proposal B and Universal Principle I are stronger than total orderings and weaker than partial orderings.

Theorem 21: Theories of grammar obeying Proposal B and Universal Principle II are stronger than total orderings and weaker than partial orderings.

In order to empirically support a $+P$ quasi ordering as preferable to or necessary over and above a total ordering (even with the addition of one of the Universal Principles to the quasi ordering), one needs to find some phenomenon which requires that rules be applied an unbounded number of times (this is why they cannot be totally ordered). It is difficult to believe that there is any such phenomenon. It would be the kind of phenomenon where, say, there is a finite but unlimited number of *a*'s before a *b* in members of *I*, and all but the last need to be deleted. The relevant $+P$ quasi ordered rule is $a \rightarrow null$ /____ab. Note that if there is an upper limit on the number of occurrences of *a*, there will be some totally ordered set of rules which will do the job. Two possible types of linguistic examples come to mind. (1) There might be phonological reasons for postulating underlying forms which have an unbounded number of contiguous consonants, or for postulating rules which have the effect of adding an unbounded number of contiguous consonants. However, for pronounciation, we want to delete all but (say) the last one. (Ringen discusses similar examples in 1976:60–61, and seems to find this extra power desirable.) (2) A syntactic rule of 'tree pruning', which says that if an NP node immediately dominates an NP node which has no sister nodes, delete it. That is, the rule:

$$[_{NP}\ [_{NP}\ X\ ]\ ] \rightarrow [_{NP}\ X\ ]$$

If there are reasons to suppose that this rule must operate an unbounded number of times on the same cycle,[34] then the added power of $+$P rules is required. However, the far-fetchedness of these examples should make one wary of postulating such rules unless the phenomenon cannot be analyzed any other way.

Theorems 22 and 23 show just how much of the 'reduction' in power mentioned in Theorems 20 and 21 is due to the fact that the rules are $+P$ type rules which have been quasi ordered and how much is due to the Universal Principles. Note that Theorems 22 and 23 guarantee that we needn't hold Position B at all, nor need we mention both of the Universal Principles.

Theorem 22: Theories of grammar obeying Proposal B and Universal Principle I are equivalent to $+P$ partially ordered rules.

Theorem 23: Theories of grammar obeying Proposal B and Universal Principle II are equivalent to $+P$ partially ordered rules.

Corollary: Given either Proposal A or Proposal B, Universal Principle I is equivalent to Universal Principle II.

9. Some concluding remarks

We take the theorems of the previous section to show that most of the discussions concerning rule ordering in the literature are drastically misguided. A number of linguistics try to show 'on empirical grounds' that one rule ordering theory is preferable to another, mostly by showing that *their* rules cannot be given any other ordering. We find it amazing that writers think they know, independently of knowing what the proper rule ordering is, what the exact rules describing any phenomenon are. After all, if one knew this (and knew the input *I*) one would automatically know the proper rule ordering. Surely, if two linguists are going to describe some phenomenon, their different attitudes on rule ordering will affect the exact form the rules take. Yet writers blithely argue for one and against another theory of rule ordering on the grounds that their(!) rules cannot give the right output unless ordered in such-and-so way, or that the opposition's rules are not 'formulated naturally'. Equally as bad, we submit, is that other writers will even bother to look for 'universal principles' whose use will allow those very rules to be subsumed under some other rule ordering. Calling one's aesthesic tastes 'linguistically significant generalizations' and then calling that an 'empirical matter' and then appealing to these supposed 'facts' as an 'empirical condition' to be imposed on the general form of rules, strikes us as an incredible case of either self-aggrandizement or self-deception. Before writers attempt to 'demonstrate that one theory of rule ordering is empirically preferable to another', or before they even treat it as an 'hypothesis' to be tested, they should investigate the logical properties of these theories, to see if it is *possible* for there to be any differences which can be empirically tested.

For example, Koutsoudas *et al.* (who seem to endorse Proposal A and Universal Principle II) say, in summary of their work:

> It was shown that, for representative facts which have been accounted for by each of the logically possible types of rule ordering relations determined by extrinsic ordering constraints, there are alternative explanations in which the order of application of rules is either entirely unrestricted, or else fully predictable from the forms of the rules by universal principles. (p. 26).

(By "extrinsic ordering constraints" they mean some total order or other of the rules.) Koutsoudas *et al.* could have saved themselves a lot of effort by noting that this result is guaranteed by Theorem 19. There is absolutely no empirical difference between their position and some theory of total ordering. Koutsoudas *et al.* also use a 'universal principle' which they call The Proper Inclusion Precedence, to help input members go through the 'randomly ordered' rules in the 'correct order' (See note 17 for further discussion.) Theorems 18–23 prove as a corollary that (given their other condition) there is never need for such other principles, because an appropriate set of rules which does not require it can always be constructed. This 'principle' (and various others suggested by members of this 'school') are simply not hypotheses in any straightforward sense of the term — they have no empirical consequences unless one already knows precisely what the rules are and what *I* is.

Chomsky (as quoted by Wall, 1971:686) once said that mathematical linguistics needs to wait for empirical linguistics to set problems for it; that it should wait

> until further empirical work on language structure manages once again to formulate concepts which are amenable to mathematical study, more intricate and complex concepts that are more well-motivated empirically.

In this paper we hope to have shown that this is not true; and that, in fact, if linguists would pay more attention to the formal nature of some of the theories they propose, they could save themselves considerable effort in futile attempts to 'empirically justify' one 'hypothesis' over another 'hypothesis', when it could have been mathematically shown that the two give rise to identical empirical claims.

Received 12 February 1980
Revised version received
2 April 1980

Department of Philosophy
The University of Alberta
Edmonton
Canada T6G 2E5

Appendix

This appendix contains proofs for most of the theorems. The proofs are (a) instructions for constructing a grammar which X-orders its rules given a grammar which Y-orders its rules (and which, given the same *I*, will produce the same *L*), or (b) features about some X-ordered grammar which no Y-ordered grammar can have (e.g. produce a language with such-and-so property). It is assumed that the reader has followed the proofs of Theorems 1 and 8; we therefore restrict our attention to showing how the rules can be 'coded up' so that the desired order for every derivations obtained.

In the discussion of constructed rules, we shall use the notation ${}^{i}p_{n}^{j}$ to indicate a rule which has [+i] in its structural description and a [+j] in its output, but which is otherwise identical with the (antecedently-specified) rule p_n. We use ${}^{i}\overset{\star}{p}_{n}^{j}$ for rules like our (b) rules in the proof of Theorem 1: where the input and output are identical (except for [+i] being replaced by

[+j]) and the structural description of it is that either the structural description or the context of p_n is *not* met. When we want to indicate some specific rule, we use its "name". Thus, for instance, in Theorem 1 we might have said: If n is the name of p_i (and p is not a first rule) construct two rules, ${}^{n-1}p_i^n$ and ${}^{n-1}\overset{\star}{\not{p}}_i^n$. (Since rules are uniquely ordered in total orders, the proof of Theorem 1 just used integers, and relied on their properties. However, in the grammars to be discussed here we need some other way of keeping track of what rule is allowed to feed what rule. Thus, our 'names'.)

Theorem 2: *Semi-partial orderings, quasi orderings, and random orderings are all at least as strong as partial orderings.*

If *P* is partially ordered, there is a unique first rule against which every member of *I* must be checked to see if it applies, before any other rule can be 'looked at'. Call this rule p_1. After p_1 has been 'looked at' for applicability, two things might happen: (a) p_1 might have applied to the first stage and is once again a candidate for application, as is p_2 (the 'second rule' of the partial ordering), or (b) p_1 might not have applied and hence we move on to p_2. We intrinsically mirror these two possibilities like this: We first construct a rule which takes any member of *I* and adds a [+1] marker to it. For the unique first rule of the partial ordering, p_1, we construct two rules — one for the case where the member of *I* (plus its [+1] marker) meets the structural description, and one for when it doesn't. The randomly ordered rule for the first case will be ${}^1p_1^{(1,2)}$ (the structural description of p_1 has [+1] added to it, and the output has both [+1] and [+2] added to it). The randomly ordered rule for the second case will be ${}^1\overset{\star}{\not{p}}_1^2$. For every rule p_i of the partial ordering we construct two rules: ${}^np_i^{(n,m+1)}$ and ${}^n\overset{\star}{\not{p}}_2^{n+1}$. All derivations start with an element of *I* and are finished when no rule applies to them. If there were *n* partially ordered rules, this will be when a stage has [+(n+1)] but not [+n] in it. We may add a final rule which replaces these by the null string, as discussed in Theorem 1. All arguments about the legitimacy of these rules and about the preservation of strong generative power which were given in Theorem 1 carry over to Theorem 2. Note that this set of rules could also be required to be quasi or semi-partially ordered and these results hold. (Note also that these rules could *not* be un-, semi, or totally ordered and still give this result.)

Theorem 3: *Random orderings, quasi orderings, unorderings and semi partial orderings are all at least as strong as semi orderings.*

A semi ordering will tell us in advance, for pairs of rules, which one of the pair must be applied first in any derivation in which they are both applied.

As remarked in note 28, there are $\frac{n^2-n}{2}$ such pairs on a set of *n* rules. In the randomly ordered grammar, have a rule which will add to any member of *I* all the markers of the form $[+\langle m, n\rangle]$ where m is the name of p_i and n is the name of p_j if and only if $\langle p_i, p_j\rangle \in \mathscr{F}'(P)$. That is, it adds $\frac{n^2-n}{2}$ markers to every member of *I*. Now, every rule of the semi ordering is a potential "first rule"; the idea is that when one is applied, then no rule which it has to come after (according to $\mathscr{F}^1(P)$ and $\mathscr{B}$) can be applied in that derivation. We mirror this by slightly altering the rules given by the semi ordering: If n is the name of p_i, then we add to the output a $[-m]$ for every m that is in a marker of the form $[+\langle m, n\rangle]$ of the string (as added by the previously-mentioned rule). In addition, we add $[-n]$ to insure that p_i will not reapply. We did this condition on the application of any rule p_i: any string (structure) to which it applies cannot contain $[-n]$ (where n is the name of p_i). This provides a randomly ordered grammar; now note that these rules could have been quasi ordered, unordered, or semi partially ordered and the result would have been the same.

Theorem 4: *Random orderings and quasi orderings are each at least as strong as unorderings.*

The conversion to random orderings is as follows. If n is the name of the rule p_i of the unordering, add the rule p_i^{-m} to the randomly ordered rules: p_j does not apply to any string (structure) with $[-n]$ in it (where n is the name of p_j). This set of randomly ordered rules could also have been required to be quasi ordered.

Theorem 5: *Random orderings and quasi orderings are each at least as strong as semi partial orderings.*

The proof of this is identical to that of Theorem 3 with the exception that p_i can be applied after (but only immediately after) p_i. For this aspect, see Theorem 6.

Theorem 6: *Random orderings are at least as strong as quasi orderings.*

If n is the name of p_i of the quasi ordering, consider the rule p_i^{-n} (this is not quite the rule to construct). The actual rule we want to construct to correspond to p_i also has the following feature (where n is the name of p_i): if the input has $[+n]$, the output continues to have $[+n]$; if the input has $[+m]$ for $m \neq n$, then the output contains $[-m]$ and $[+n]$. The condition on application of any rule p_i is that it cannot apply to a string (structure) with $[-n]$ in it, if n is p_i's name.

Theorem 7: *Total orderings are at least as strong as simultaneous application.*

Given that our simultaneously applied rules are not "contradictory", they could be given *any* ordering and the same input would result in the same output, but for two obstacles: necessary environments might be destroyed and unwanted ones might be created. We eliminate these possibilities by means of 'tags' and creation of a few new rules. Starting with some ordering of the original rules, we 'tag' the output of each rule with a special symbol — different symbols for the different rules. Already this guarantees that the second difficulty, that of creating unwanted environments, is eliminated (since the structural descriptions and contexts of none of the rules is met by the output of any of them now). So all we need to worry about is when necessary structure is destroyed. We correct this by adding an appropriate rule. Such rules are constructed whenever the structural descriptions and contexts of two rules overlap one another. Consider these three rules to be applied simultaneously:

1. $a \rightarrow b \quad /c\underline{\quad\quad}d$
2. $c \rightarrow e \quad /\underline{\quad\quad}a$
3. $b \rightarrow f \quad /\underline{\quad\quad}d$

We can replace these three rules by the following four (where the superscripts serve as the 'tags' mentioned above).

1.* $a \rightarrow b^1 \quad /c\underline{\quad\quad}d$
2.* $c \rightarrow e^2 \quad /\underline{\quad\quad}a$
3.* $b \rightarrow f^3 \quad /\underline{\quad\quad}d$
4.* $c \rightarrow e^2 \quad /\underline{\quad\quad}b^1$

These four rules, when applied in the order 1–4, will give exactly the output that the original three did when applied simultaneously, except for occurences of the 'tags'. If desired, one can add the appropriate deletion rules to the end of the totally ordered list of rules, following the method outlined in Theorem 1. It should be noted that this total ordering gives exactly the same statements about relationships between members of I and L as did the simultaneously applied set.

Theorem 9: *Semi partial orderings are stronger than semi orderings.*

At the first stage of any derivation of a semi ordering, a finite number of rules are candidates for application. At any other stage of that derivation, a lesser number of rules are candidates for application than were at the immediately previous stage. Hence, all derivations end after a finite number of stages. This is not the case with a semi partial ordering, which allows 'infinite' derivations by means of rules like a → ab. Hence languages can be produced by semi partial orderings which cannot be produced by any semi ordering. Together with Theorem 3, this proves Theorem 9.

Theorem 10: *Quasi orderings are stronger than unorderings.*

The proof of this is identical to that of Theorem 9, except that Theorem 4 is used.

Theorem 11: *Random orderings are stronger than quasi orderings.*

Suppose $I = \{a\}$ and $P = \{a \rightarrow ab, a \rightarrow ca\}$. If P is randomly ordered, every derivation will be 'infinitely long'. Note that at every stage of derivation, each of the two rules is a candidate for application. Therefore, there are $2^{\aleph_0}$ derivations. Now consider any finite set P' of rules which are quasi ordered. We can indeed construct these rules so that every derivation is 'infinitely long', but we cannot construct so many derivations. For, given any $m > 0$, the number of derivations that have $m + 1$ stages is no greater than $(m \cdot n)$ (where n is the number of rules in the set). Thus the number of derivations with an infinite number of stages is no greater than $n \cdot \aleph_0 = \aleph_0$, which is less than $2^{\aleph_0}$. Hence there are languages which random orderings can produce that quasi orderings can't. together with Theorem 6, this proves Theorem 11.

We now turn our attention to proofs which do not show that there is any difference in weak generative capacity of the rule orderings involved, but rather attend to other aspects of strong generative power, in particular output relatedness. By this we understand: given an I, a P, and a rule ordering theory of type Y, two members of the output language L are *P-related according to theory Y* if and only if ${}^{P}D^{i}$ and ${}^{P}D^{j}$ are identical for their first k stages. (A special case is where they only have the first stage — a member of I — in common.) We shall show that there are members of L which, given I, can be P^1-related according to theory Y but which cannot be P^2-related according to theory X, for any P^2, without also producing outputs which are not members of L.

Theorem 12: *Semi partial orderings are stronger than partial orderings.*

In any partial ordering there is a unique 'first' rule. Semi partial orderings do not have this feature; therefore, there can be ${}^{P}D^{h}$, ${}^{P}D^{i}$, and ${}^{P}D^{j}$ of a semi partial ordering of P which are identical in their first stage but all other stages are different. This can happen with no partial ordering; for, either this first stage meets the structural description of the 'first' rule or it doesn't. If the first stage does meet the structural description of the 'first' rule, then all these derivations will apply it and hence they all have identical second stages. If it doesn't meet the structural description of this rule, than the 'second' rule can be treated as if it were the 'first' rule. Together with Theorem 2, this proves Theorem 12.

Theorem 13: *Unorderings are stronger than semi orderings.*

To prove this, we give a partially specific example. Let p_1 be the rule a → ab. In an unordering this rule cannot be reapplied once it has been applied. Note that unorderings allow us to use p_1 and some p_2 in different orders in different derivations. Now, while this cannot be done in a semi ordering, we can do something having a similar effect; we can give a different set of rules, say p'_1, p'_2, p'_3, such that one derivation of a semi ordering uses the rules and order $\langle p'_1, p'_2\rangle$ to achieve the effect of the unordering's using $\langle p_1, p_2\rangle$, and another derivation of the semi ordering uses the rules and order $\langle p'_2, p'_3\rangle$ in order to achieve the effect of the unordering's $\langle p_2, p_3\rangle$. Let us now consider the case where in the output language *L* of the unordering, the $\langle p_1, p_2\rangle$ and $\langle p_2, p_1\rangle$ derivations yield the same string. Now note that the semi ordering is unable to mirror this. It requires that $\langle p'_1, p'_2\rangle$ be a member of its *L* and that $\langle p'_2, p'_3\rangle$ be a member of its *L*. But given our example of p_1 (it creates its own input), and the definition of semi orderings (they are connected, so that either $\langle p'_1, p'_3\rangle$ or $\langle p'_3, p'_1\rangle$ is a member of $\mathscr{F}^1$), it follows that the semi ordering will generate a sentence ruled out by the unordering. Such an example can always be constructed for any semi ordering which is attempting to match an unordering.

Theorem 14: *Quasi orderings are stronger than semi partial orderings.*

The proof of this theorem parallels that of Theorem 13, except that rules are allowed to apply immediately after themselves. So the precise string that would be generated in addition to the members of *L* would be different. (In other words, the analogue of the new derivation to $\langle p'_3, p'_1\rangle$ would already be present in the quasi ordering. One needs a similar example with one more intervening rule.)

Theorem 15: *Total ordering and simultaneous application are equivalent.*

If the simultaneously applied rules are not 'contradictory', then they can be totally ordered following the method of Theorem 7. And again, so long as the rules are not 'contradictory', the totally ordered rules will not say to do 'contradictory things' to any part of the input — such as generate intermediate representations which get collapsed onto other ones from other parts of the input — any such totally ordered set of rules can simply be required to be simultaneously applied without even changing the rules. (As an example, consider the case of only two rules which are totally ordered. If neither rule 'feeds' the other in such a way that an intermediate representation will 'get lost', it is obvious that this direction is proved. Yet this is precisely what happens in the case that the rules are not 'contradictory'.)[35]

Theorem 16: *Semi orderings and partial orderings are not comparable.*

(a) As remarked in the proof of Theorem 9, all derivations of a semi ordering are finite. Partial orderings, however, can generate 'infinite' derivations.
(b) In a semi ordering of n rules, there might be as many as n derivations whose first members are identical but all other stages are different. This cannot be done in a partial ordering; for even if one has sufficient rules so that there are exactly n distinct derivations whose first members are identical, to this first stage only one rule can apply and so it is not the case that all these derivations are identical *only* in their first stage.

Theorem 17: *Unorderings and semi partial orderings are not comparable.*

The proof of this follows that of Theorem 16.
Theorem 18: *Theories of grammar obeying Proposal A and Universal Principle I are equivalent to total orderings.*

(a) That total orderings can be converted into *P sets of unordered rules follows from Theorem 1. (The set of rules actually constructed there obeyed the Ringen condition and could be required to be unordered.
(b) *P sets of rules contain 'copies' of some rules. For simplicity, assume that each rule has the same number of 'copies'; that there are n rules in *P and there are m 'distinct' ones of them (call the set containing only the 'distinct' rules P.) A total ordering which has the same power as *P if the rules obey the Ringen condition is: First, list the rules P in any order, adding $[+1]$ to the output of each. Then list them again in any order, adding $[+1]$ to their structural descriptions and $[+2]$ to their output, and so on n times ($=$ number of members of *P). This set of rules will contain $m \cdot n$ rules, which is finite. Now, if the rules obey Ringen conditions, the first stage of any derivation will meet the structural description of at most one of the first m rules. If it is altered, it will get a $[+1]$ in its structure. Again, this structure will meet the structural description of at most one of the rules numbered between $m+1$ and $2m$; and so on. Note that every derivation here is identical with a derivation in the Ringen *P rules.

Theorem 19: *Theories of grammar obeying Proposal A and Universal Principle II are equivalent to total orderings.*

(a) That total orderings can be converted to *P rules obeying the Derwing/KSN principle follows from Theorem 1.
(b) The other direction follows from the remarks in part (b) of the proof of Theorem 18, together with Theorem 15. The conversion goes like this. The first stage of any derivation is some member of I, each succeeding stage was to come from simultaneously applying all the rules in P (the set from

which *P was constructed). The simultaneous application of this set is continued for some finite number of times (guaranteed by the fact that they are *P rules). Each individual time the rules are simultaneously applied can be totally ordered, as Theorem 7 showed; and this preserves strong generative capacity, as Theorem 15 shows. This set of rules is repeated for each stage of any derivation (we use the method of Theorem 18 to keep these sets 'separated'). By the argument of part (b) of Theorem 18, we know that we need *n* sets of these rules, so the total number of rules in the total ordering will be finite.

Theorem 20: *Theories of grammar obeying Proposal B and Universal Principle I are stronger than total orderings and weaker than partial orderings.*

We divide the proof into four parts.

(a) That they are at least as strong as total orderings follows from Theorem 1.

(b) That they are stronger than total orderings follows from (a) and the comments of part (b) of the proof of Theorem 21.

(c) The proof that partial orderings are at least as strong as these is similar to that of part (b) of Theorem 18. The differences are (1) every rule of +P is listed in each of the 'groups' (rather than merely listing the 'distinct' ones), (2) in order to preserve antisymmetry, we add to the output of every rule of each group another marker [+n], where n is the name of that rule. (These markers are different from, and in addition to, the markers we used in Theorem 18.) We furthermore add that if the input to a rule whose name is m has a [+n] marker, and $m \neq n$, then [+n] becomes [−n]. Finally a condition on application of any rule whose name is n is that its input cannot contain [−n]. Note that since the set of rules we here have constructed is partially ordered, if a rule has applied, it is a candidate to reapply immediately; but once some other rule has applied it is no longer a candidate. We take it to be part of the Ringen condition (when we talk of quasi orders) that if a stage is the result of applying rule p_i, and this stage continues to meet the structural description of p_i, then it will not also meet the structural description of any other rule. Note that every derivation legitimatized by this partial ordering was also legitimatized by the +P quasi ordering obeying Ringen conditions, and conversely.

(d) Partial orderings allow 'infinite derivations', which no *P unordered set of rules can.

Theorem 21: *Theories of grammar obeying Proposal B and Universal Principle II are stronger than total orderings and weaker than partial orderings.*

(a) Follows from Theorem 1.

(b) As discussed in the text, +P quasi ordered rules allow a rule to apply an unbounded number of times to its own output. Since they are +P rules, this will not lead to 'infinite derivations'; and so given a particular *I*, for every set of quasi ordered +P rules, there is *some* set of totally ordered rules (say P′) yielding the same *L* from this *I*. But any constructed set of rules will be inadequate because a different *I* will yield distinct *L*'s from the quasi ordered +P set and the P′ totally ordered set. Hence it is not true that they are equivalent in the sense of always yielding the same *L* from the same *I*.

(c)–(d) See discussion in Theorem 20.

Theorem 22: Theories of grammar obeying Proposal B and Universal Principle I are equivalent to +P partially ordered rules.

Follows from part (c) of Theorem 20. (Note that the effect of using +P rules instead of just any legitimate rule is what caused the difference in Theorem 20.)

Theorem 23: Theories of grammar obeying Proposal B and Universal Principle II are equivalent to +P partially ordered rules.

Follows from part (c) of Theorem 21.

Notes

[*] Catherine M. Fletcher was essential in the preparation of this paper. It was she who first interested me in the topic; and we have had numerous conversations concerning especially the matters of Sections 5–9. In fact, earlier versions of this paper were presented as jointly authored. Stan Peters has generously shared his time and his knowledge of mathematical linguistics with us. A considerable part of Section 6 is due to discussions with him. For their comments on earlier versions, we would like to thank Noam Chomsky, Kit Fine, John Heintz, Gerald Gazdar, C. G. Morgan, Geoff Pullum, Helmut Schnelle, Len Schubert, and Steve Thomason. Earlier versions of this paper were read to the Canadian Linguistics Association, 1975; Dept. Linguistics, Univ. Washington; and the Society for Exact Philosophy, 1977.

[1] We include the $B \neq e$ condition even though it is not strictly necessary. If one allows $e \in Vt$, then for any language which is generated by a grammar which is otherwise context free except that it allows rules of the form $A \rightarrow e$, there is a grammar that is otherwise context free except that it has one rule $S \rightarrow e$, where S is the start symbol and S does not appear on the right of any production. (For proof see Hopcroft and Ullman, 1968:62–63.) We wish, however, to keep context free languages a proper subset of context sensitive languages, and this is not true using the revised definition of context free grammars with our definition of context sensitive grammars. (The reason is that context sensitive languages are not closed under arbitrary substitution mappings, but they are if the language is *e*-free. See Hopcroft and Ullman, 1968:124ff.).

[2] Current linguistic theory, at least of the sort of interest here, has evolved from Chomsky (1965) through Chomsky's (1973, 1975) 'trace theory' to the more current Chomsky (1977). The mechanisms used in these later works can all be defined in the (1965)

framework. The 'constraints' introduced in the later works will be discussed below in Section 4. Other competing linguistic theories have been developed which do not seem to be compatible with Chomsky (1965): generative semantics and Montague grammars in particular. For discussion see Newmeyer (1980) or Pelletier (1977). For the view that they are equivalent see Cooper and Parsons (1976).

3. For more detailed statements of the restrictions imposed on transformational grammars, see Peters and Ritchie (1973a). It should be noted that many, more modern, treatments have more than just *S* as a cyclic node — usually *NP* is also added.
4. The few studies that have implications for strong generative capacity are Chomsky (1955: Chapters 6–7), Chomsky (1963), Chomsky and Schützenberger (1963), Crespi-Reghizzi (1971), and Langendoen (1979).
5. There is, of course, lexical ambiguity — a word or other lexical item having more than one sense. We are not here concerned with that kind of ambiguity at all.
6. We thank Gerald Gazdar for this example, although he doesn't approve of the analysis of it as having one surface structure. See Gazdar (1979a, 1979b).
7. For reasons that we do not fully understand, Peters and Ritchie (1973a) state their results in terms of weak generative capacity *after* elimination of the bracketization, leaving themselves open to the charge of eliminating what is linguistically interesting, namely what they call (1973a:69) the strong generative capacity of transformational grammars (where this is taken to be such matters as surface structural ambiguity, constituent structure, etc.). The results could have been stated as well in terms of what *we* are calling weak generative capacity, which includes all these notions that can be defined on surface structure. Peters (1969) points out that the strong generative capacity of transformational grammars is equivalent to turning machines so long as the functions defining the relevant features of strong generative capacity are recursive. He does not there note that the definition of the function which removes brackets to form the terminal string in Peters and Ritchie (1973a) guarantees that (surface structural) ambiguity is a recursive function.
8. Langendoen's (1979) 'direct generation' of labelled phrase markers (his 'bracket diagrams') appears to neglect this. His method will give a final phrase marker for every sentence, but in the cases like (4) above, it will not suffice for disambiguation. He nonetheless calls this mirroring of structure in the final phrase marker 'the strong generative capacity' of a grammar.
9. Some linguists insist on including as empirical reflexes such features as simplicity or elegance of rules and the ability to conflate apparently distinct concepts by one symbol in the statement of rules. (The latter is often called a case of 'finding linguistically significant generalizations'). Following what we think is more responsible linguistics, we treat these kinds of things as 'matters of aesthetics'.
10. The result had already been proved in Kimball (1967), but in a somewhat more complicated way.
11. Of course, this proves that there is at least one such grammar. There may be infinitely many of them, so it is not sufficient merely to disallow the deletion transformation mentioned. (Such a mistaken suggestion can be found in Skousen, 1972, under the guise 'empirically based rule').
12. The earliest of these sorts of criticisms appears to be Putnam (1961), but they have become quite common in the intervening years.
13. More detailed discussions can be found in Derwing (1973), Seidenberg *et al.* (1977), Levelt (1974).
14. Too many linguists, it seems, rely on this kind of 'internal-to-the-theory' evidence. Such reliance is akin to claiming that there is a sense in which baseball outfielders have

'internalized' Newtonian Laws of Motion on the grounds that their position after a baseball is hit can be computed 'in an economical manner' from facts like the speed of the pitch, the angle of the bat, etc. plus the Newtonian Laws. Any such inferences are ill-founded unless there is a method for independent psychological checking.

15. See Watt (1970), Fodor *et al.* (1975), Marslen-Wilson (1975) for detailed documentation of the lack of correlation.
16. Peters and Ritchie only consider *S* to be a cyclic node.
17. Some examples of such general principles are Koutsoudas *et al.* 'proper inclusion precedence principle', Ringen 'obligatory precedence principle', and so on. The kind of thing claimed (for the former principle) is that if an input simultaneously meets the structural description of two distinct rules, the rule that has a structural description which properly includes the other will apply first.

    (This 'proper inclusion precedence' is unfortunately ambiguous. Following Pullum (1979) we take 'has a structural description which properly includes the other's' to mean 'applies to a proper subset of structures of the other'. Consider Pullum's (1979:50) two rules EXTRAP and ITDEL which have the following structural descriptions:

    EXTRAP: $X\ [_{NP}\ \textit{it}\ \ [_{S}\ \textit{that/for}\ Y]_{S}\ ]_{NP}\ Z$

    ITDEL: $X\ [_{NP}\ \textit{it}\ \ [_{S}\ Y\ ]_{S}\ ]_{NP}\ Z$

    It is claimed that EXTRAP applies first, when the principle is followed, since the structural description of EXTRAP has the symbol *that /for* which ITDEL doesn't have. Note, however, that in another sense of 'proper inclusion', the structural description of ITDEL properly includes that of EXTRAP, since the former is satisfied by any input which the latter is satisfied by.)
18. The original idea behind intrinsically ordered rules was that one needn't state an order when (say) one rule creates a structure that the other will use. We broaden this somewhat to mean any theory that does not claim an explicit order on the application of the rules, regardless of whether the rules end up 'ordering themselves' by always creating the input for another.
19. However, as we shall show below, there are many theories of grammar other than this 'no-order' theory that have this feature.
20. A further flaw is Pullum's italicized claim that linear ordering theories allow the *maximum* number of grammars. Below we consider theories which are similar to ones presented in the literature under the heading 'local ordering' that allow many more numbers of grammars. See note 28.
21. Some arguments for rule orderings are simply invalid, even on their own terms. See the sections in Chapter I of Pullum (1979) on 'the fallacy of ignoring cyclicity', 'the strict order fallacy', 'the fallacy of insufficiency', and so on. See also Koutsoudas (1978).
22. Levine (1976).
23. Their official statement (see, e.g. 1971:483) makes them be ordered and in (1971) they are applied in that order to give the result. In (1973), only one transformation was required when the base language was context sensitive, and the rule was applied 'iteratively' (see below for definition).
24. Salomaa (1971) proved the result for the Ginsburg and Partee formulation of transformational grammar.
25. We are aware of the uneasiness that might be caused by our allowing the possibility of non-terminating derivations. The 'pure' theories here considered do allow this possibility. In Section 8 we consider possible remedies for this possibility. We should

also mention here the possibility of 'vacuous' application of rules, i.e. where some input meets the structural description of a rule but the application of the rule yields no change in the input. In a totally ordered theory this is harmless, since the input will simply go on to the next rule; but in some of the theories we are considering here, where a rule can be reapplied, this could raise severe difficulties. We therefore adopt the convention that there is some mechanism for discovering this and preventing a rule from reapplying if there was no change from the time it previously was applied. (See Ringen, 1976: PRINCIPLE V).

26. It is to be understood by these conditions that 'the relation imposed on *P*' is imposed on the transitive closure of *P*, and the restrictions mentioned are that this transitive closure be either asymmetric or antisymmetric. It is further understood that if there is more than one rule in *P*, every rule occurs as either the first or the last element of $\mathscr{F}$(P) — i.e. the transitive closure is connected.

27. Besides obvious cases for context free rules, such as *a* → *b* and *a* → *c* not being available for simultaneous application to any string containing *a*, there are less obvious cases. Rules such as *ab* → *ba*/d__c and *bc* → *cb*/a__e cannot be simultaneously applied to *dabce*. In the case of transformations (letting *X* and *Y* be variables), X *a* Y → 2 1 3 and X*b* Y → 2 1 3 cannot be simultaneously applied to the string *cabd*. Another example is a rule which is to de-stress syllables that are surrounded on each side by stressed syllables. It is difficult to see how to formulate this rule to apply to sequences of four or more consecutive stressed syllables (or even what the output should be for ordered rules).

Some rules obviously do not dictate these 'contradictory changes', and it is to sets of these non-contradictory rules we refer when discussing simultaneous application. Our condition below is intended to pick these sets out. Although we call this phenomenon 'contradictory rules', it is not so much that the rules are contradictory as it is that perfectly normal-appearing rules might have input to which they cannot coherently be simultaneously applied.

28. As we mentioned above in notes 19–20, Pullum's claim that a theory of grammar invoking random ordering is preferable because it 'allows the minimum number of grammars consistent with it', viz, one, and his claim that a theory of grammar invoking linear (= our total) ordering is worst because it allows the 'maximum' number of grammars consistent with it, viz. n! (when *n* is the number of rules), are deficient. In this regard, a theory of grammar which says that in any grammar the rules are to be quasi ordered, or a theory that says they are to be unordered, or a theory that says they are to be simultaneously applied, also have the property that they 'allow only one grammar'. Pullum's claim about total orderings is mirrored by us as follows: if there are *n* rules in *P*, there are n! distinct ways to state $\mathscr{F}^1$(P) or $\mathscr{F}^2$(P), and since in total and partial orderings there are unique 'first rules', it follows that a theory of grammar which says that every grammar must be totally (or partially) ordered can give rise to n! distinct grammars from a given set of *n* rules. With respect to semi and semi-partial orderings of *n* rules, there are also n! grammars (since there are that many $\mathscr{F}^1$(P) or $\mathscr{F}^2$(P)). However, a given derivation is allowed to 'start anywhere in the ordering', and so there is a sense in which these theories of grammars 'invoke a wider class of grammars' than total or partial orderings — namely that for every n! derivations allowed by the latter types, the former types allow $2^{(\frac{n^2-n}{2})}$ derivations. (In this regard compare the slightly misleading remarks in Wasow, 1975:376 footnote, and see also Pullum's remarks in his 1979b:180–181.) It is not as obvious to us as it is to Pullum that a semi partial ordering invoking six rules is three times as bad an explanation of some phenomenon as a total ordering invoking eight rules, nor 32,768 times as bad as a quasi ordering invoking 100 rules. We think that a more central notion of 'stronger theory' is to be put in terms of

strong generative capacity, since it is here that we capture all empirical differences between theories.

29. The 'features' mentioned in this rule — *high*, *back*, *round*, etc. — are binary: a linguistic form is marked + or − for a particular feature. If a rule mentions a feature, then the form must have the value for that feature mentioned in the rule. This example is a complex rule with deletion, alternation, and addition of features. The change indicated by the rule takes place in a context.
30. 'A' here is a variable. The statement of the condition is in full accord with 'standard' linguistic practice. The careful reader of Langendoen (1979:footnote 6) might also note that the statement of the condition can be put as part of the structural description, when we use his 'direct generation' method. As he notes, this has the effect of 'removing from linguistic theory the artificial distinction between conditions on form and conditions on applicability' which standard linguistic theory seems to embody.
31. Ringen (1976) seems to be the only exception.
32. Thus a particular version of this proposal will be our unorderings, where the 'finite number' is one.
33. They may seem different, but reflection will show that Koutsoudas *et al.* 'apply the rules sequentially but in no particular order until no more rules apply, but whenever the structural description of more than one rule is met apply them simultaneously' is the same as Derwing's 'apply all rules simultaneously over and over until no more apply'.
34. So this extra power would not be necessary for

$$[_S\,[_S\,X]] \rightarrow [_S\,X]$$

This deletion need be stated but once, and it would work on each *S* cycle. Similarly, the extra power would not be necessary if *NP* were a cyclic node.

35. This is not to say that any total ordering can be converted into a simultaneously applied set of non-contradictory rules. It is rather to say that, if the rules are non-contradictory, then a total ordering can be converted into a simultaneously applied set of rules by using this method. For the actual use we shall make of Theorem 15, namely in Theorems 19 and 21, our weaker proof is exactly what is needed.

References

Anderson, S. (1969). West scandanavian vowel systems and the ordering of phonological rules. Ph.D. dissertation (distributed by Indiana Univerisity Linguisitcs Club), Indiana University.

Bach, E. (1974). *Syntactic Theory*. New York: Holt, Rinehart and Winston.

Burt, M. K. (1971). *From Deep to Surface Structure: An Introduction to Transformational Syntax*. New York: Harper Row.

Chomsky, N. (1955). *The Logical Structure of Linguistic Theory*. New York: Plenum Press, 1975.

—(1957). *Syntactic Structures*. The Hague: Mouton.

—(1962). The logical basis of linguistic theory (914–978) and Discussion (978–1008). *Proceedings Ninth International Congress of Linguists.*

—(1963). Formal properties of grammars. In R. Luce, R. Bush and E. Galanter (eds), *Handbook of Mathematical Psychology, Volume II*, 323–418. New York: Wiley.

—(1965). *Aspects of the Theory of Syntax*. Cambridge, Mass.: MIT Press.

—(1971). Deep structure, surface structure, and semantic interpretation. In D. Steinberg and L. Jokobovits (eds), *Semantics*, 183–216. Cambridge: Cambridge University Press.

—(1973). Conditions on transformations. In S. Anderson and P. Kiparsky (eds.), *Festschbrift for Morris Halle*, 215–226. New York: Holt, Rinehart and Winston.

—(1975). *Reflections on Language*. New York: Pantheon Books.

—(1977). On WH-movement. In A. Akmajian, P. Culicover and T. Wasow (eds), *Formal Syntax*, 71–132. New York: Academic Press.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. New York: Harper and Row.

Chomsky, N. and Lasnik, H. (1977). Filters and control. *Linguistic Inquiry* 8, 425–504.

Chomsky, N. and Schützenberger, M.-P. (1963). The algebraic theory of context free languages. In P. Braffort and D. Hirschberg (eds), *Computer Programming and Formal Systems*, 118–161. Amsterdam: North-Holland.

Cooper, R. and Parsons, T. (1976). Montague grammar, generative semantics, and interpretative semantics. In B. H. Partee (ed.), *Montague Grammar*, 311–362. New York: Academic Press.

Crespi-Reghizzi, S. (1971). Reduction of enumeration in grammar acquisition. *Second International Conference on Artificial Intellegience*, 546–552.

Derwing, B. (1973). *Transformational Grammar as a Theory of Language Acquisition*. Cambridge: Cambridge University Press.

Dinnsen, D. (1972). *General Constraints on Phonological Rules* Ph.D. dissertation, (distributed by Indiana University Linguistics Club).

—(1974). Constraints on global rules in phonology. *Language* 50, 29–51.

Fiengo, R. (1977). On trace theory. *Linguistic Inquiry* 8, 35–62.

Fodor, J., Bever, T. and Garrett, M. (1975). *The Psychology of Language: An Introduction to Psycholinguistics and Generative Grammar*. New York: McGraw-Hill.

Gazdar, G. (1979a). Constituent structures. (unpublished manuscript.)

—(1979b). English as a context free language. (Unpublished manuscript.)

Ginsburg, S. and Partee, B. H. (1969). A mathematical model of transformational grammars. *Information and Control* 15, 297–334.

Gross, M. and Lentin, A. (1970). *Introduction to Formal Grammars*. Berlin: Springer Verlag.

Hastings, A. J. (1976). On the obligatory-optional principle. In A. Koutsoudas (ed.), *The Application and Ordering of Grammatical Rules*, 294–319. The Hague: Mouton.

Hopcroft, J. and Ullman, J. (1968). *Formal Languages and Their Relation to Automata*. Reading, Mass.: Addison-Wesley Publishing Co.

Johnson, C. D. (1972). *Formal Aspects of Phonological Description* (= Volume 3 of *Monographs on Linguistic Analysis*). The Hague: Mouton.

Joshi, A. K. and Levy, L. S. (1977). Constraints on structural descriptions: local transformations. *SIAM Journal Computing* 6, 272–284.

Kenstowicz, M. and Kisseberth, C. (1973). The multiple application problem in phonology. In C. Kisseberth (ed.), *Studies in Generative Phonology*, 1–12. Edmonton: Linguistic Research Inc.

Kimball, J. (1967). Predicates definable over transformational derivations by intersection with regular languages. *Information and Control* 11, 177–195.

Koutsoudas, A. (ed.) (1976a). *The Application and Ordering of Grammatical Rules*. The Hague: Mouton.

—(1976b). Unordered rule hypotheses. In A. Koutsoudas (ed.), *The Application and Ordering of Grammatical Rules*, 1–21. The Hague: Mouton.

—(1978). The question of rule ordering: some common fallacies, (distributed by Indiana University Linguistics Club).

Koutsoudas, A., Sanders, G. and Noll, C. (1974). The application of phonological rules. *Language* 50, 1–28.

Langendoen, T. (1979). On the assignment of constituent structures to the sentences generated by a transformational grammar. *City University of New York Forum.*

Lapointe, S. (1977). Recursiveness and deletion. *Linguistic Analysis* 3, 227–266.

Lasnik, H. and Kupin, J. (1977). A. restrictive theory of transformational grammar. *Theoretical Linguistics* 4, 173–196.

Levelt, W. J. M. (1974). *Formal Grammars in Linguistics and Psycholinguistics*, 3 Vols. The Hague: Mouton.

Levine, A. (1976). Why argue about rule ordering? *Linguistic Analysis* 2, 115–124.

Lightfoot, D. (1977). On traces and conditions on rules. In A. Akmajian, P. Culicover and T. Wasow (eds), *Formal Syntax*. New York: Academic Press.

Marslen-Wilson, W. (1975). The limited compatibility of linguistic and perceptual explanations. CLS *Parasession on Functionalism*, 409–420.

McCawley, J. (1968). Concerning the base component of a transformational grammar. *Foundations of Language* 4, 243–269.

Newmeyer, F. J. (1980). *A History of Transformational Grammar*. New York: Academic Press.

Newton, B. (1971). Ordering paradoxes in phonology. *Journal of Linguistics* 7, 31–53.

Palacas, A. (1971). Iteration of rules vs. infinite schemata in phonology. (Paper presented at Linguistic Society of America meeting.)

Pelletier, F. J. (1977). ([How/Why]) does linguistics matter to philosophy? *Southern Journal Philosophy* 15, 393–426.

Pelletier, F. J. and Fletcher, C. M. (in preparation). Optional rules.

Perlmutter, D. (1973). Evidence for a post-cycle in syntax (Paper presented at Linguistic Society of America meeting.)

Peters, S. (1969). A note on the universal base hypothesis. *Journal of Linguistics* 5, 150–152.

—(1970). Why there are many universal bases. *Papers in Linguistics* 2, 27–43.

—(1973). On restricting deletion transformations. In M. Gross, M. Halle and M.-P. Schützenberger (eds), *The Formal Analysis of Natural Language*, 372–384. The Hague: Mouton.

Peters, S. and Ritchie, R. (1971). On restricting the base component of transformational grammars. *Information and Control* 18, 483–501.

—(1973a). On the generative power of transformational grammars. *Information Sciences* 6, 49–83.

—(1973b). Non-filtering and local-filtering transformational grammars. In J. Hintikka, J. Moravcsik and P. Suppes (eds), *Approaches to Natural Language*, 180–194. Dordrecht: Reidel.

—(1973c). Context sensitive immediate constituent analysis: context free languages revisited. *Mathematical Systems Theory* 6, 324–333.

Pullum, G. (1979). *Rule Interaction and the Organization of a Grammar*. New York: Garland.

—(1979b). Review of Koutsoudas (1976a). *Journal Linguistics* 15, 179–187.

Putnam, H. (1961). Some issues in the theory of grammar. Reprinted in G. Harmon (ed.), *On Noam Chomsky* (1974), 80–103. Garden City: Anchor Books.

Ringen, C. (1972). On arguments for rule ordering. *Foundations of Language* 8, 266–273.

—(1976). Vacuous application, iterative application, reapplication and the unordered rule hypothesis. In A. Koutsoudas (ed.), *The Application and Ordering of Grammatical Rules* 55–75, The Hague: Mouton.

Salomaa, A. (1971). The generative capacity of the transformational grammars of Ginsburg and Partee. *Information and Control* 18, 227–232.

Seidenberg, M. and Tanenhaus, M. (1977). Psychological constraints on grammars: on trying to put the *real* back into 'psychological reality'. *CLS* 13.

Skousen, R. (1972). Empirical Restrictions on the Power of Transformational Grammar. *Papers in Linguistics* 5, 250–269.

—(1975). *Substantive Evidence in Phonology: The Evidence from Finnish and French*. The Hague: Mouton.

Soames, S. (1974). Rule ordering, obligatory transformations, and derivational constraints. *Theoretical Linguistics* 1, 116–138.

Wall, R. (1971). Mathematical Linguistics. In W. O. Dingwall (ed.), *A Survey of Linguistic Science*, 683–712. College Park: University of Maryland Press.

—(1972). *Introduction to Mathematical Linguistics*. Englewood Cliffs: Prentice-Hall.

Wasow, T. (1975). Anaphoric pronouns and bound variables. *Language* 51, 368–383.

—(1978). On constraining the class of transformational languages. *Synthese* 39, 81–104.

Watt, W. C. (1970). On two hypotheses concerning psycholinguistics. In J. R. Hayes (ed.), *Cognition and the Development of Language*, 137–220. New York: Wiley.