On the Logic of Variable Rules
Author(s): Paul Kay and Chad K. McDaniel
Published by: Cambridge University Press
Stable URL: http://www.jstor.org/stable/4167070
Accessed: 30/06/2013 15:34

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.
On the logic of variable rules

PAUL KAY AND CHAD K. McDANIEL

University of California, Berkeley

I. INTRODUCTION

The variable rule method was introduced by William Labov in 1969 and has been refined and extended by a number of scholars, principally Cedergren & Sankoff (1974). It is the only method of analysis so far proposed that can by its nature deal with all variable linguistic data, and it is the only method for dealing with such data that self-consciously accords to the tacit ability of each member of the speech community a direct reflection of the quantitative variation existing in the speech community and observed by the linguist (G. Sankoff 1974). The method is becoming increasingly popular because of the theoretical contributions of its proponents and also because it is so easy to use. There is now a complete variable rule analysis package available from David Sankoff through which anyone may run his observed data, and there is every reason to believe that increasing numbers of investigators will be inclined to subject observed data on variation to an automated analysis with such an excellent pedigree.

It is with the underlying logic, the covert and semi-covert assumptions, of the variable rule method that this paper is concerned. We do not believe this logic has been explicitly stated in the literature to date. We will conclude that each of the versions of the variable rule methodology as currently formulated entails hypotheses to which the linguist will not (or should not) wish to commit himself.

We will further conclude that variable ‘rules’ are best conceived not as linguistic rules in anything like the usual sense of that term. Rather the variable rule methodology provides a family of mathematical models that dictate probabilistic constraints on possible distributions of linguistic variants in a speech community and a statistical method for deriving measures of the relative importance of different determining factors from a matrix of observed linguistic variation. The

[1] We have freely utilized comments on an earlier draft of this paper by Frank Anshen, Derek Bickerton, Ralph Fasold, Don Forman, William Labov, and an anonymous reader for Language in Society. We would like to express our appreciation to these people and particularly to David Sankoff, whose highly detailed and insightful comments provided much of whatever value the present essay contains. We have not accepted all the good advice offered us.

[2] The major competing analytical device is the implicational scale (DeCamp 1971; Bickerton 1971; Bailey 1973). Implicational scale analysis can of course succeed only when the data scale, and it is not the case that all variable data of interest scale implicationally (see Kay 1978). D. Sankoff & Cedergren (1971) have also attempted to apply more general multidimensional scaling methods to linguistic data.
PAUL KAY AND CHAD K. MCDANIEL

variable rule method thus provides a battery of related statistical models and accompanying data reduction techniques for the descriptive analysis of a certain kind of linguistic corpus.

A variable rule analysis is closely related both mathematically and philosophically to an application of the statistical technique called analysis of variance. A typical application of analysis of variance would be to a situation in which one wanted to know which determining variables, such as soil type, available moisture, sunlight, fertilizer type, etc., were most important in producing a heavy yield of, say, corn. Analysis of variance is a technique that reduces a set of data giving, for example, the amount of corn produced under all combinations of determining factors to a set of measures of the relative importance of the determining factors (and their interactions). The results of such an analysis do not in and of themselves constitute a substantive theory of how corn plants grow, although perusal of such analyses may be valuable to the plant physiologist or economic botanist in attempting to form such a theory. It will be the conclusion of this paper that variable rule analyses should be viewed in much the same light, not as providing direct theoretical insight into the substantive processes that produce linguistic variation, but as a statistical tool that may be of considerable heuristic value to those searching to discover and understand such processes.

Our criticisms of the variable rule methodology are not philosophical but technical. We are very much in favor of quantitative analysis of quantitative linguistic data and do not object to variable rules either because they violate linguistic orthodoxy or because they violate certain a priori beliefs about human psychology. Objections of both kinds have been made and it may be well to consider and reject a couple of them at the outset so that these a priori criticisms of the variable rule method may be contrasted with those we wish to make.

One popular objection to variable rules is that, contrary to some implications in Labov's writings, they represent not an amendment to traditional generative grammar but a departure so radical that it cannot be conceived within the orthodox framework. On this view, variable 'rules' are not rules at all in the sense defined in generative grammar; hence their formulation represents not an extension of generative theory but a de novo departure which leads to a conceptual muddle in so far as its proponents think they are working within the generative framework. Briefly the argument runs as follows. A generative grammar is a device that performs several functions, the chief one of which is to specify the membership of a set, in particular the set of sentences that constitutes a natural language such as English. (Other functions include providing structural descriptions, and so on.) That is, a language is conceived as a set of sentences and the primary task—one might say the defining task—of a grammar is to sort potential sentences into those that are in the language and those that are not. For example, in Syntactic Structures, Chomsky (1957: 30) presents for expository purposes a simple, non-natural language consisting of all sequences of
ON THE LOGIC OF VARIABLE RULES

some number n a's followed by n b's: \{ab, aabb, aaabbb, \ldots\}, and provides for it a grammar that may be rendered

(1) a. \(S \rightarrow ab\)

b. \(ab \rightarrow aSb\)

We may note that rule (1)b is optional. Let us imagine that we add a third, optional, rule, a context-sensitive rule that converts the last \(a\) to \(c\).

(2) \(a \rightarrow c/b\)

With rule (2) added, our grammar of course is the grammar of a different 'language', in particular the language \(\{ab, cb, aabb, aaabbb, aacbbb, \ldots\}\). The point is that optional rules like (1)b and (2) are quite at home in generative grammar; in fact they are its main stock in trade.

Some variationists have thought that variable rules are just optional rules gone a little more precise. If an optional rule says, 'before \(b\) realize \(a\) as \(c\) sometimes', it does not seem at first blush a major departure to make this a little more precise by saying, 'before \(b\) realize \(a\) as \(c\) 67\% of the time'. The relation of the variable rule to the optional rule appears to be the relatively minor one of specifying an actual numerical value of a parameter whose presence was implied anyway by the 'sometimes' clause of the optional rule. The perceived relation is roughly that between the ordinary sentences, 'Some nights John snores', and 'Two out of three nights John snores'. (Cf. Labov 1966: 761): 'I do not regard these methods or this formal treatment as radical revisions of generative grammar or phonology.'

But as many workers in the generative tradition have realized – some perhaps more consciously than others – variable rules do constitute a radical departure. The reason for this is that optionality in an optional rule does not mean that the rule is sometimes applied. Viewed from the point of view of the variable rule theorist, obligatory ('categorical') rules are those that are reflected in the production of all relevant tokens, optional rules are those that are reflected in the production of some relevant tokens and variable rules are those that are reflected in the production of a specified proportion of the relevant tokens (in a very large sample). But from the generative point of view, this is quite wrong because grammar deals only with types, never with tokens. In the generative framework, an optional rule is just one that adds a new class of sentences (seen as linguistic types) to the set of sentences that is generated by the grammar. ('Generate' here has, of course, nothing to do with the real time production of actual utterances (= sentence tokens) by real people.) In the example above, rule (2) simply adds to the set of sentences generated by rules (1)a,b the set \(\{cb, accb, aacbbb, \ldots\}\). In generative theory, the starting point is the set of sentences (i.e., sentences as types) that constitutes the 'language'. The frequency with which a sentence is produced as an utterance (token) is simply irrelevant. Hence a 'rule' which is
concerned with predicting token frequencies is not a rule of (generative) grammar.

This argument is, we feel, unexceptionable, and Labov was wrong in 1969 in
supposing he was not proposing a radical departure from generative grammar. It
is, however, another matter to say that this argument stands as an indictment of
variable rules as a useful tool in studying language, because such an indictment is
viable only to the extent that the orthodox generative view of language is the
only profitable view. In particular, this argument only indicts variable rules if it
is correct that linguistic theory must equal grammatical theory and grammatical
theory must be based on the definition of a language as a set of sentences. The
truth of neither of the latter statements is self-evident, and although much useful
work has been done taking these as foundational assumptions, it appears that a
lot of other useful work has been and will continue to be done within approaches
to language that do not accept these statements as first principles. In short, to
show that variable rules are not, strictly speaking, generative rules is not to
demonstrate they are of no use in furthering our understanding of language.

In addition to relatively formally stated challenges to variable rules, such as
that sketched above, there have been several of a more substantive kind. Many
of these involve conflicting evaluations of the empirical evidence for the existence
of inherent variability, as, for example, the continuing debate represented in
G. Sankoff (1973), Cedergren & Sankoff (1974), Bickerton (1973), D. Sankoff &
Rousseau (1974) and McDaniel (1975). It is not our intention to review all of
these issues, nor in general would we presume to attempt to adjudicate the multi-
dimensional conflict between the wave theory and variable rule approaches to the
study of linguistic variation and change in process. There is, however, one further
a priori criticism of the variable rule method that has been so frequently put
forward. It should be countered before we proceed to a more technical critique
of the variable rule method. The essence of this argument is the belief that the
human mind cannot handle probabilities, at least with respect to linguistic
behavior. The clearest statement of the argument is perhaps that expressed by
Bickerton:

An obligatory rule says: 'When you recognize environment $S$, use feature $Y$' –
a straightforward enough operation. A variable rule, however, says 'When you
recognize environment $X$, use feature $Y Z$% of the time.' $Z$ does not, of course,
represent a precise figure. Labov does not envisage that the behavior of a
member of a rule-sharing group will necessarily be isomorphic with that of all
or even any of the other members, though it is true that he does not expect it to
vary much: 'It is unlikely that it will be important for us to know that the
copula is deleted 82% of the time by speaker A and 79% of the time by speaker
B' (1969: 740). No doubt he would also be prepared to admit that A's 82%,
may itself be only an average of scores ranging between say 77% and 87%.
However, in order that the average for his group should remain constant, the
variation of the individual must be confined within a relatively narrow range. What keep his percentages within those limits? And how can it keep within them unless something, somewhere, is counting environments and keeping a running score of percentages? Nor is it merely tokens of a few environments per variable item that will have to be counted. Labov envisages rules in which up to six features may exercise constraints; for each of such rules, $6 \times 6 = 36$ different environments would have to be counted and all their percentages maintained in the correct order of magnitude. Nor is this all. For, since the group figure is the crucial one, and since (within the approved limits) individual scores will vary around it, each individual must— if that group figure is to be maintained— keep track, not merely of his own environments and percentages, but also of those produced by all other members of his group; in other words, speaker B must continually be saying to himself things like: ‘Good Lord! A’s percentage of contractions in the environment $+_V+_+$ has fallen to 77! I’ll have to step up mine to— let’s see: A’s production of this environment-type stands to mine in the ratio 65:35 over the last 100 token-occurrences, so I’d better compensate by shooting up to... what? About 86%?’ And, to crown it all, he must not only be able to perform all these highly sophisticated calculations— he must also (since the rules apply to ‘single’ as well as ‘group’ styles) somehow continue to do so even in the physical absence of all other group-members (Bickerton 1971: 460–61, emphasis in original). Despite its superficial plausibility, this argument is seriously flawed. There is no more reason to suppose that the attribution of probabilities to the mental-neural process resulting in variable speech behavior requires that the speaking organism possess a mechanism to keep track of prior performance than there is to suppose that the attribution of probabilities to a physical process resulting in, say, roughly equal numbers of the variants ‘head’ and ‘tail’ requires that a coin be endowed with such a counting and memory mechanism. Hence, there is no problem in counting variants that the speaker has not heard because there is no problem in counting variants at all. To suppose that a device which sequentially produces variant outcomes in conformity with a probabilistic model is perfectly equipped with a feedback mechanism that keeps track of prior performance is to misunderstand the theory of probability and its applications. On the contrary, none of the physical devices, such as those used in gambling games, which provide the paradigmatic physical models of the mathematical theory of probability are possessed of such capabilities. For the law of large numbers to be an accurate model of the behavior of a device does not entail that the device keep a record of its performance. Opponents of variable rules have offered no arguments

---

PAUL KAY AND CHAD K. MCDANIEL

why this should be true of humans when it is not true of physical mechanisms whose behavior conforms to probabilistic laws.

To be sure, probabilities are high-powered, abstract mathematical objects and to attribute the ability to deal with probabilities to speakers is to make a strong claim. But strong claims about the nature of the mathematical complexity of the mental-neural abilities that underlie language use are hardly foreign to linguists. As mathematical objects probabilities don’t seem to us any more high-powered than the familiar elements of generative grammars, pan-lectic grammars, or what have you.

Secondly, there is empirical evidence that humans do have the ability to learn and apply probabilities quite unconsciously and naturally. The phenomenon of probability learning, including the special case of ‘probability matching’, is well attested in experimental psychology.

Grant, Hake, and Hornsby (1951)... showed that when subjects are asked to guess which of two events will occur, and when these two events occur with unequal probabilities, then the subjects will guess each event in approximately the same proportion as the event actually occurs. This result, which has led to the ‘probability-matching hypothesis,’ has generated a considerable amount of research to determine why the phenomenon happens, and the conditions under which it does not. The concern about the phenomenon is that such behavior does not conform to a particular type of optimum game strategy, since the maximum number of correct guesses will always occur when the subject guesses the more probable event 100 percent of the time. In other words, it is assumed that subjects are in fact trying to maximize their ‘hits’ (Garner 1962: 84).

Although it is well attested that under certain conditions humans reliably produce probabilistic behavior in response to probabilistic stimuli, there is no general agreement regarding the theoretical explanation of this fact. According to Sternberg (1963: 62):

The validity of this finding and the particular conditions that lead to it have been the subjects of considerable controversy. Partial bibliographies may be found in Edwards (1956, 1961), Estes (1962), and Feldman & Newell (1961). The reader should also consult Restle (1961, Chapter 6) and, for work with several nonhuman species, the papers of Bush & Wilson (1956) and of Bitterman and his colleagues (e.g. Behrend & Bitterman, 1961). The general conclusions to be drawn are that the phenomenon does not occur under all conditions or for all species, that when it seems to occur the response probability may deviate slightly but systematically from the outcome probability [i.e. the probability of the stimulus], that matching may characterize a group average although it occurs for only a few of the individuals within the group, and that an asymptote may not have been reached in many experiments.
ON THE LOGIC OF VARIABLE RULES

Since human subjects have been shown capable of learning and using probabilistic information unconsciously in at least some non-linguistic tasks, the burden of proof would seem to fall on those who wish to claim that humans cannot learn and apply probabilistic information in the domain of linguistic behavior. So far no experimental or observational evidence for this claim has come to our attention. We conclude that neither probability theory nor the weight of existing psychological evidence condemn things like variable rules as preposterous hypotheses a priori.

2. VARIABLE RULES: THE ADDITIVE MODEL

The underlying assumptions of variable rules may perhaps be best understood in the context of the historical development of this methodology. This history has two major phases, marked by two principal methodological articles: Labov (1969) and Cedergren & Sankoff (1974). (In singling out these two articles we intend no slight to other important contributions and contributors to the variable rule methodology.) In the first article Labov proposed what has come to be known as the additive model, and in the second Cedergren and Sankoff suggested that the additive model be supplemented or replaced by one or another version of what they have called the multiplicative model. We will consider in turn the empirical motivations and formal character of each of these models.

In presenting his data on token frequencies of contraction and deletion of the forms is and are in the speech of whites and blacks in New York City, Labov noted strong statistical regularities in the frequencies with which certain non-obligatory phonological rules appear to operate depending upon the linguistic environment. For example, in the texts that Labov collected, the rule that contracts ‘John is going’ to ‘John’s going’ is virtually obligatory for all speakers when the subject NP is a pronoun (‘He is going’ → ‘He’s going’). When the preceding NP is not a pronoun, contraction is more likely to occur when the subject NP ends in a vowel than when it ends in a consonant (‘Martha’s going’ is a more likely occurrence than ‘Robert’s going’). With respect to the grammatical category of the constituent that follows the (possibly) contracting is, contraction is most likely to occur preceding a future in gonna (John’s gonna go), less likely before any other sort of VP (‘John’s eating’), even less likely before a predicate adjective or locational phrase (‘John’s happy’, ‘John’s in the bathroom’), and least likely before an NP (‘John’s a good man’). These observations regarding relative frequencies of utterance tokens hold up reliably across four black teenage and preteenage groups, one group of black adults and one group of white teenagers. The general pattern of these data is summarized in Fig. 1.

When stable relative frequencies are observed in nature, the usual model to apply is a probabilistic one, and this is what Labov did. The above account of the patterning in the data is considerably simplified and we will fictionalize the data
Further in representing Labov’s theoretical formulation. Our purpose here is not
to characterize Labov’s substantive findings of New York speech but to examine
his method of analysis. Labov formalized these observations in what is now called
an additive variable rule. Recall that contraction is more likely to occur following
a vowel than following a consonant and more likely to occur preceding a VP
(including the future auxiliary gonna) than before any other sort of following
environment (NP, PA, or Loc). Some fictitious frequencies and proportions
conforming to these empirically accurate ordinal statements are given in Table 1.
(Our motive for using fictitious frequencies in this example will become clear
later.)

Table 1. Fictitious Data Corresponding to Labov’s (1969) Observations on
Relative Frequencies of Application of the Rule Contracting is According to
Preceding and Following Linguistic Environment

<table>
<thead>
<tr>
<th>Following verb</th>
<th>Preceding vowel</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>99/110</td>
<td>91/130</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>109/218</td>
<td>36/120</td>
<td></td>
</tr>
</tbody>
</table>

a. 'observed' frequencies

In Table 1a, the entry 99/110 in the upper left hand cell means that in the
texts being analyzed 110 instances of some form of is preceded by a vowel and
followed by a verb were observed of which 99 were contracted ('s) forms. The
corresponding entry in Table 1b expresses this ratio as a decimal.

The various environments (+[V-], −[V−], +[−Vb], −[−Vb]) appear to
influence differentially the likelihood that the rule of contraction will apply. As
we have noted, a preceding vowel favors contraction over a preceding consonant
(0.9 > 0.7; 0.5 > 0.3) and a following verb is more favorable to contraction than any
ON THE LOGIC OF VARIABLE RULES

other following environment (0.9 > 0.5; 0.7 > 0.3). Can these ordinal observations be more precisely quantified? An obvious way to try to do this is to discover a procedure which (a) solves a data table such as Table 1 for a set of numerical weights corresponding to each environment, i.e. each row and column, and (b) provides a rule for algebraically combining the row and column weights corresponding to a particular cell so as to produce a predicted frequency of occurrence (= 'probability of rule application') for that cell which corresponds closely with the actual frequency observed in that cell. If the predicted frequencies match closely the observed frequencies, the procedure is vindicated. The rule for algebraic combination of weights involved in such a procedure is called a variable rule. Types of variable rule, 'models' in the terminology of Cedergren & Sankoff (1974), are named after the form of the equation used to combine the weights, which are often called 'variable constraints'. If only addition and subtraction are involved in the equation, the model is called 'additive'; if multiplication and division are also involved, it is called 'multiplicative'. So far there has been a single additive model proposed and two distinct multiplicative models.

The additive model is the one originally proposed by Labov in 1969, and we take it up first. All variable rules posit – in addition to the 'variable constraints', which are the weights corresponding to the different linguistic environments – an 'input' constraint, which may be thought of as representing the speaker's inherent proclivity to apply the rule regardless of linguistic environment. When more than a single type of speaker is being modeled, types of speaker (e.g. classes, sexes, races, or combinations of these or other social categories) may be differentiated by their input constraints. The variable (= linguistic) constraints are generally assumed to be shared tacitly throughout the speech community with individual speakers or homogeneous social groups differing from each other only on the input constraint. This assumption lies back of the common notion (disputed in Kay 1978) that variable rules provide the basis for a community grammar that simultaneously captures what is shared and what varies in the speech community: the linguistic constraints are uniform in the community and only input constraints vary from speaker to speaker. This assumption leads to empirical problems which will be discussed later. For the moment it suffices that we bear in mind that in the variable rule methodology the grammars of different speakers (or groups of speakers) in a given speech community may be distinguished, with respect to a particular variable rule, only by having distinct numerical values for the input constraint, never by having distinct numerical weights attached to a linguistic environment.

We are now prepared to develop the additive variable rule model, using for illustration the hypothetical data of Table 1. We will have five constraint parameters to estimate: the input constraint plus a variable constraint corresponding to each row and each column. Once we have obtained the parameters, there will be three of them involved in the calculation of each predicted cell frequency
(1) \( V \rightarrow \emptyset \left\langle \frac{+V}{-V} \right\rangle a \quad \# \quad \left\langle \frac{-V}{+V} \right\rangle b \)  

where the variable parts of the rule are enclosed in angled brackets. (Ignoring the variable parts, rule (3) says that a vowel is deleted when it is the first segment of a one- or two-segment verb 'word, the second segment of which, if any, is a consonant.) One may note that each angled bracket contains what has been called a 'family' of mutually exclusive contains. Only one constraint from each family – and exactly one constraint from each family – is involved in a particular application of the rule. For example, to each cell of Table 1 there corresponds a selection of one constraint from family \( a \), corresponding to a column of Table 1, and one constraint from family \( b \), corresponding to a row of Table 1. In the additive model the sum of the constraints in a family is conventionally set at zero (though other conventions would be possible; for example, we could require that the mean value be \( \frac{1}{2} \)). Thus if the family contains two constraints, as has been the case in the vast majority of empirical applications of this model, the two constraints have the same absolute value and opposite sign; if a weight of \( 0.4 \) is assigned to the presence of a preceding vowel, a weight of \( -0.4 \) is automatically assigned to presence of a preceding consonant or pause (i.e., absence of a preceding vowel).  

In order to express the various forms of variable rules in common terms, it is convenient to introduce at this point another bit of notation. Each dimension in a data table corresponds to a variable constraint family. For example, in the two-dimensional tables considered in this paper, the rows collectively represent one constraint family and the columns another. The data covered by a given variable rule will contain as many such tables as there are groups of speakers distinguished by their input parameters. We first designate the constraint families \( a, b, \ldots, n \). In the present example we may take the 'a' constraint family to correspond to the preceding environment \( +[V-] \) vs. \( -[V-] \) and the 'b' constraint family to correspond to the following environment \( +[-Vb] \) vs. \( -[-Vb] \). Now when speaking of a particular cell in the table we refer to the relevant constraint from constraint family \( a \) as \( p_a \), the relevant constraint from constraint family \( b \) as \( p_b \), and so on. In the present example, when we

---

[4] There is no theoretical limitation on the number of constraint families, nor is it necessary, or even usual, that only one constraint family appear at a given linear position. (We are using 'constraint family' in discussing the additive model as roughly equivalent to what Labov (1969) calls a 'variable constraint' or 'variable feature'.)
ON THE LOGIC OF VARIABLE RULES

speak of the lower left-hand cell \( p_a \) is the parameter associated with the feature value \(+[V-] \) and \( p_b \) is the parameter associated with the feature value \([-\neg Vb] \). In this notation, introduced by Cedergren & Sankoff (1974), expressions like \( 'p_a' \) are systematically ambiguous, but the context, if properly specified, always removes the ambiguity. When a particular cell is specified by the context, \( 'p_a' \) refers to one particular variable constraint as specified above. When a general rule is given, however, without specification of a particular cell – as in the equations specifying the various variable rule models – an expression like \( 'p_a' \) really stands for a variable ranging over all the cells and may be understood to mean 'the constraint from the "a" family relevant to the cell under consideration'.

The symbol \( 'p' \) is of course chosen to be mnemonic for 'probability', the notion being that a variable rule is concerned with the probability of a certain step in a derivation occurring. The idea is that there is a weight associated with each variable feature and these are combined by the variable rule into the various probabilities of application in all the possible environments defined by the cells of the table. The decision to interpret variable rules in terms of probabilities has consequences that will be examined presently.

What is the most natural way to assign and combine the numerical weights? Labov's not unreasonable approach boils down to the following. The simplest way to combine weights is to add them. Labov (1969: 738) considered the input constraint to be inherently inhibitory so that the basic tendency of the rule to apply (regardless of conditioning environment) was given by the expression \( 1 - p_0 \). There seems to be no good reason to choose \( 1 - p_0 \) over the simpler expression \( p_0 \) to represent this quantity and so, following a suggestion of D. Sankoff, we adopt the latter course. This basic tendency is augmented in favorable environments and reduced in unfavorable ones by the relevant weights. Thus the rule begins with the input constraint \( p_0 \) and adds or subtracts the weights corresponding to the relevant environments, yielding as the equation for the additive model

\[
(4) \quad \text{Additive model} \quad p = p_0 + p_a + \ldots + p_n
\]

where \( 0 < p_0 \leq 1 \) and \( -1 \leq p_i \leq 1 \) for \( p_i \in \{p_a, \ldots, p_n\} \). (The where-clause ensures that the \( 'p_0' \) term is always positive and allows for either positive or negative variable constraints; when some weight \( p_i \) is negative, the resulting \( + p_i \) term in (4) will of course be negative also.)

We would like now to see how this model applies to data like those in Table 1. In order to do this, however, we must adopt another simplifying fiction. If a probabilistic model is the correct model for a body of observed data, the theory of probability tells us that the observed statistics descriptive of the data will only rarely correspond exactly to the underlying probabilities. For example, if we toss a fair coin ten times, the most likely single outcome is five heads and five tails, but if we do not obtain exactly five heads and five tails we would be foolish to reject
the hypothesis that the coin is fair, since the mathematical expectation of precisely this observed outcome, given a fair coin, is a little less than one fourth. The same principle applies to a variable rule. The combination of weights in a variable rule yields a probability of application of the rule. We cannot expect the observed frequencies to correspond exactly to the underlying probabilities even if the model is correct. What is done in practice is to apply a statistical technique to the observed data which is called maximum likelihood estimation. This technique yields for a given body of data and a given variable rule model – in the present instance the additive model – the set of underlying constraint values most likely to have produced the observed result, assuming that the additive variable rule model is in fact the correct model for the process underlying the observed data. Returning to the coin example, if five heads and five tails are observed, the maximum likelihood estimate of the underlying probability of a head is one half. If six heads and four tails are observed, the maximum likelihood estimate for the underlying probability of a head is 0.6; and so on. But if the coin is in fact fair (probability of heads = ½) the probability of getting exactly six heads (and four tails) in ten tosses is slightly more than one fifth, which is appreciable when we recall that with the same fair coin the probability of five heads (and five tails) is only about one fourth.

We therefore adopt the fiction for the purpose of exposition that the observed frequencies of Table 1 have happened to correspond to the single most likely outcome of the model. It must be emphasized that this assumption is made for expository purposes only and that it is not legitimate to make it in practice. (In practice one uses maximum likelihood estimation.) As we have seen, in the coin example the single most likely outcome occurs less than one fourth of the time. Another way to put this is that we are acting as if our observed frequencies are precisely the expected frequencies generated by the model, an event that would rarely occur in practice. Thus, the sample computations given below really relate the predicted frequencies of the model to the underlying constraint values.

For the (fictitious) data summarized in Table 1 (and recalling that we are treating observed frequencies as predicted frequencies) equation (4) reduces to

\[ p = p_0 + p_a + p_b \]

which in fact is a summary of four simultaneous linear equations.\(^5\)

\[ \begin{align*}
(6) &. \quad 0.9 = p_0 + p_+V + p_V + verb \\
&. \quad 0.7 = p_0 + p_-V + p_V + verb \\
&. \quad 0.5 = p_0 + p_+V + p_-_verb \\
&. \quad 0.3 = p_0 + p_-V + p_V + verb
\end{align*} \]

\[^5\] Hereafter we abbreviate \( \pm [V\_] \) as \( \pm V \) and \( \pm [\_\verb] \) as \( \pm \verb \).
ON THE LOGIC OF VARIABLE RULES

We recall that by convention

(7) a. \( P_{-v} = 1 - P_{+v} \)
   b. \( P_{-verb} = 1 - P_{+verb} \)

So we really have four equations in just three unknowns, \( P_0, P_{+v}, P_{+verb} \), since \( P_{-v} \) and \( P_{-verb} \) are automatically determined by \( P_{+v} \) and \( P_{+verb} \), respectively in (7). That is, equations (6) may be rewritten as

(8) a. \( 0.9 = P_0 + P_{v} + P_{verb} \)
   b. \( 0.7 = P_0 - P_{v} + P_{verb} \)
   c. \( 0.5 = P_0 + P_{v} - P_{verb} \)
   d. \( 0.3 = P_0 - P_{v} - P_{verb} \)

The reader may recall that a system of four linear equations in three unknowns does not in general yield a unique solution, or for that matter any solution, since any three of the equations determines a unique solution and in general, different selections of three equations produce different solutions. However, we chose the fictitious data in Table 1 so that the resulting equations would have a unique solution; namely \( P_0 = 0.6, P_{verb} = 0.2, P_v = 0.1 \). That is, when one substitutes these values into equations (8a-d), each of which presents an analysis of a cell frequency in Table 1, they become identities. Variable rules are really systems of equations which relate the numerical values attached to each individual constraint on rule application to the probability that the rule will be applied under each possible combination of environmental constraints.

The (fictitious) observed frequencies we selected for Table 1 have the property that there is a constant difference between the corresponding entries in each row and also a constant difference between the corresponding entries in each column. In particular the constant row difference is 0.4 and the column difference is 0.2, as may be verified in the table. The general property that row and column differences are each homogeneous characterizes any set of expected frequencies generated by an additive variable rule regardless of the numerical values of the underlying constraint weights, although as the values of the constraint weights change, the numerical values of the differences will change. As we have observed at some length, observed frequencies will not usually correspond exactly to predicted frequencies even when the rule is correct. On the other hand, the more the observed frequencies depart from any possible set of expected frequencies, the less likely it is that the additive rule is the correct model. This observation gives us a handy, though vague, quick check on the applicability of the additive model. If the observed frequencies yield nearly homogeneous row and column differences, we know that the data can be well fit by the additive model; to the extent that

163
the row and column differences are heterogeneous we know that the data cannot be fit well by the additive model.6

Our reason for selecting hypothetical data to exemplify the internal mathematics of the rule was to have a perfect fit of the data to the model in the sense that the (fictitious) observed frequencies and the expected frequencies arrived at by the maximum likelihood estimation procedure were identical. Let us now consider the actual results Labov found for the preceding vowel and following verb constraints, which are the constraints to which he gives major emphasis in his study (Labov 1969: 749; Figs 12 & 13). Table 2 shows the observed frequencies of operation of contraction and deletion of is among adolescent black New Yorkers as constrained by preceding vowel and following verb in Labov's (1969) study. It is clear from the table that the additive model fits the deletion data well and the contraction data poorly in that the column and row differences are nearly homogeneous in the former case and quite heterogeneous in the latter.

The generalization of this quick test for the additive model in any data table is that for each constraint family X, for any particular selection of constraints from the other constraint families the difference between the Yes (X) and No (∅) cells should remain constant. For example, if there are three binary constraint

[6] Although we shall take no further note of it in this paper, each of the variable rule models to be developed below places comparable, though distinct, constraints on possible relative frequencies. These are most easily summarized in tabular form. The homogeneity of row-and-column differences in the additive model means that the entries in a four-fold table of expected frequencies in this model must display the relations shown in the following table:

Constraint family a
Constraint family b
<table>
<thead>
<tr>
<th>a1</th>
<th>a2</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>b + c - a</td>
</tr>
</tbody>
</table>

The comparable tables for the three other variable rule models to be discussed below are:

Multicative application probabilities model

<table>
<thead>
<tr>
<th>a1</th>
<th>a2</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>bc/a</td>
</tr>
</tbody>
</table>

Multicative non-application probabilities model

<table>
<thead>
<tr>
<th>a1</th>
<th>a2</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>bc/(a + (1-b)(1-c)(1-a))</td>
</tr>
</tbody>
</table>

Logistic model

164
ON THE LOGIC OF VARIABLE RULES

TABLE 2. Observed Token Frequencies for Contraction and Deletion of is in Labov’s (1969) N.Y. Black Teenage Data, Showing that Row Differences and Column Differences are Each Heterogeneous for Contraction and Homogeneous for Deletion

<table>
<thead>
<tr>
<th>V→</th>
<th>+Vb</th>
<th>Column differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+0.90</td>
<td>0.77</td>
</tr>
<tr>
<td>Row differences</td>
<td>-0.75</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Observed frequencies for contraction rule in Labov’s data

<table>
<thead>
<tr>
<th>V→</th>
<th>+Vb</th>
<th>Column differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+0.78</td>
<td>0.43</td>
</tr>
<tr>
<td>Row differences</td>
<td>-0.95</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Observed frequencies for deletion rule in Labov’s data

families A, B, C, there are eight cells in the data table as represented in Table 3. The following equalities should then be satisfied:

(9) \[ ABC - \overline{ABC} = \overline{ABC} - ABC = ABC - \overline{ABC} = \overline{ABC} - \overline{AB} \overline{C} \]

\[ ABC - A\overline{BC} = ABC - \overline{ABC} = A\overline{BC} - \overline{ABC} = \overline{ABC} - ABC \]

\[ ABC - \overline{AB}C = A\overline{BC} - \overline{ABC} = \overline{ABC} - \overline{AB}C = \overline{ABC} - \overline{AB}C \]

where, for example, \( ABC \) is the observed frequency in the cell representing Yes on features A and B and No on feature C.

TABLE 3.

<table>
<thead>
<tr>
<th>A</th>
<th>( \overline{A} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>ABC</td>
</tr>
<tr>
<td></td>
<td>ABC</td>
</tr>
<tr>
<td>( \overline{B} )</td>
<td>ABC</td>
</tr>
<tr>
<td></td>
<td>ABC</td>
</tr>
</tbody>
</table>

We have seen that in the additive model a numerical weight between zero and unity inclusive is assigned to the input constraint and a weight between \(-1\) and \(+1\) to each variable constraint; the probability of application of the rule is, in each environment, the sum of the relevant weights. There is, however, more to the model than this. Additional restrictions must be placed on the underlying
constraint weights to assure that the resulting sum, the probability of application, falls within the required interval \([0, 1]\). In particular, equations (10, 11, 12) jointly express a sufficient condition that the probability of application comes out in the interval \([0, 1]\). Equations (11) and (12) are necessary conditions, but equation (10) is not. That is, every set of weights yielding a meaningful value for \(p\) satisfies (11) and (12), while some satisfactory sets of weights do not satisfy (10).

(10) \(0 \leq p_0 \leq 0.5\)

(11) \(p_0 \geq p_i\) for \(p_i \in \{p_a, p_b, \ldots, p_n\}\)

(12) If the constraint families are arranged in descending order of their positive (or absolute) values so that \(p_1 \geq \ldots \geq p_n\), then

\[
p_i \geq \sum_{i=1}^{n} p_j (i, j \in \{1, 2, \ldots, n\})
\]

The meaning of conditions (10), (11), and (12) is easiest explained in inverse order. Condition (12) says that the effect of an arbitrary constraint must be at least as great as the combined effect of all the less important constraints. An example where this condition is not met will show why this must be so. Suppose \(p_0 = 0.5, p_a = \pm 0.4, p_b = \pm 0.3\). Conditions (10) and (11) are met, but (12) is not since the sum of constraints \(p_a\) and \(p_b\) exceeds constraint \(p_0\). Consider the environment in which the negative values of constraints \(a\) and \(b\) are in effect. In this case the variable rule (4) has the values

(13) \(p = p_0 + p_a + p_b\)

\[
p = 0.5 + (-0.4) + (-0.3)
\]

\[
p = -0.2
\]

Of course it is meaningless to say that the probability of a rule's applying is \(-0.2\).

Condition (11) says that the input constraint must have the largest value and condition (10) says that this value must be between zero and 0.5 inclusive. Let us consider examples where (10) and (12) are satisfied but not (11) and where (11) and (12) are satisfied but not (10). Example (14) satisfies (10) and (12) but not (11). Assume that negative values of constraints \(a\) and \(b\) are in effect.

(14) \(p_0 = 0.1, p_a = \pm 0.5, p_b = \pm 0.3\)

\[
p = 0.1 - 0.5 - 0.3
\]

\[
p = -0.7
\]

One cannot have a probability of \(-0.7\).

Examples (15)a,b satisfy conditions (11) and (12) but not (10). Assume in (15)a that the positive values of the variable constraints are operative.
ON THE LOGIC OF VARIABLE RULES

(15) a. \( p_0 = 0.6, \quad p_a = \pm 0.4, \quad p_b = \pm 0.1 \)

\[ p = 0.6 + 0.4 + 0.1 \]

\( p = 1.1 \)

(15) b. \( p = 0.6, \quad p_a = \pm 0.2, \quad p_b = \pm 0.1 \)

A probability of 1.1 is impossible, whereas in (15)b, \( 0 \leq p \leq 1 \) for every selection of constraints. We have seen that in order for the probability of application to come out between zero and unity, conditions (11) and (12) must each be satisfied. That is, each is a necessary condition on the additive model's yielding a 'probability' of application within the interval [0, 1]. Conditions (10), (11), and (12) jointly constitute a sufficient condition that \( 0 \leq p \leq 1 \), but we will not show that here.

Conditions (11) and (12) are apparently what Labov intends by his numbered statement (95) (Labov 1969: 741), judging from the surrounding text. Labov does not appear to be aware that conditions (11) and (12) do not jointly provide a sufficient condition for \( 0 \leq p \leq 1 \). A perhaps more serious failing in Labov's presentation of these matters is that his text is worded in such a way that some readers may have concluded that his 'postulate of geometric ordering' (Labov 1969: 741) is empirically motivated. Labov, referring to the predicted cell frequencies as 'cross products', says

The constraints of a preceding pronoun and a following noun phrase are not equivalent: they are ordered in relation to each other. This ordering is most apparent in the relationships of the cross-products, where one feature is favorable and the other unfavorable. If no statements could be made about the relationships of such cross-products, then we would have a very weak type of ordering; a strong statement would be that all of the cross-products are strictly ordered. We can formalize such a strong postulate of geometric ordering as follows:

(95) If \( X_1, X_2, \ldots, X_n \) are variable constraints upon a rule \( r \), then for any given \( X_1, X_2, \ldots, X_{i-1}, \phi_r(X_i) > \phi(\sim X_i) \).

In other words, each constraint in the hierarchy outweighs the effects of all constraints below it. If we take sentences with \( \alpha \) fixed, then any subset of these with \( \beta \) as plus will show the rule applying in a higher proportion of cases than any subset with \( \beta \) as minus. The cross-product with \( \beta \) as plus and \( \gamma, \delta \ldots \) as all minus will still show a higher value of \( \phi \) than the cross-product with \( \beta \) as minus and all lower constraints as plus (Labov 1969: 740–1).

It is not important for the reader of this paper to be concerned with the precise meaning of the symbols in Labov's statement (95) as they are not fully specified in Labov's original paper. The words, 'each constraint in the hierarchy outweighs the effects of all the constraints below it', do, however, seem to have the

167
intent of our condition (12), and elsewhere in Labov’s text it is made clear that he intends condition (11) to hold also. The important thing to realize about all this is that Labov’s ‘postulate of geometric ordering’, represented more clearly in our (11) and (12), is not a theoretically motivated substantive claim but rather a purely mathematical exigency of the chosen form of the additive model of variable rules. If the assumed underlying constraints are combined according to the rule (4), then conditions (11) and (12) have to be imposed in order to avoid application probabilities less than zero or greater than unity.

What happens, one may ask, when the observed data do not yield underlying parameters conforming to conditions (11) and (12), which is most of the time? One approach is to truncate, by which is meant simply to rename any ‘probability’ that exceeds unity ‘unity’ and to rename any ‘probability’ that comes out less than zero ‘zero’ (Cedergren & Sankoff 1974: 337). This solves the problem only in the weak sense of providing a means of getting a predicted frequency between zero and unity for every cell (Cedergren & Sankoff 1974: 337), but since the truncation method amounts to an arbitrary transformation of the data to render it mathematically tractable, it cannot be considered a solution of the problem in any interesting sense.

These drawbacks to the additive model are recognized by Cedergren and Sankoff and they cite them as motivation for developing their multiplicative models. There is an additional drawback to the additive model that Cedergren and Sankoff do not cite explicitly. They mention that ‘the additive model has been criticized on the grounds that it posits a numerical computation facility as part of competence’ (Cedergren & Sankoff 1974: 337), and reject this criticism, correctly we think, on the basis that it is substantially the same as the sort of a prioristic psychological speculation that we rejected above. But there is a related although distinct drawback to the additive model that Cedergren and Sankoff don’t point out, and their failure to do so is curious in that (a) this drawback is quite damning and (b) the multiplicative models are not subject to it.

We have seen that in the additive model, linguistically unmotivated numerical assumptions must be made before we can interpret the p on the left side of defining equation (4) as a probability. But at least this interpretation, although strained in these ways, is conceivable. If we consider, however, the underlying constraints posited in this model, we see immediately that they cannot possibly be probabilities, since half of them are negative. In fact, the numbers attached to the underlying constraints in the additive model have no interpretation at all. They are ‘weights’ in an abstract model, but they have no claimed empirical interpretation. They are just some numbers that one calculates on the basis of a data table using an arbitrary rule of calculation to produce a resulting number which is then interpreted as the probability that a speaker will apply the rule. But this is not really a probabilistic model, because the basic parameters in the model are not probabilities.
3. THE MULTIPLICATIVE MODELS

Cedergren & Sankoff (1974) were certainly aware of the fundamental flaw in the additive model just discussed, though they indicate that awareness only implicitly by being clear about the probabilistic character of the multiplicative models.

A constraint family in the multiplicative model corresponds to a set of mutually exclusive environmental features and is interpreted as a set of probabilities corresponding to a set of mutually exclusive events.\(^7\) In our running example, there are two constraint families, one corresponding to the phonological environment preceding *is* and one to the grammatical environment following *is*.\(^8\) The first constraint family contains the constraints 'probability of contracting *is* when preceded by a vowel' symbolized \(p_v\) and 'probability of contracting *is* when preceded by a non-vowel' symbolized \(p\sim_v\). The second constraint family contains the constraint 'probability of contracting *is* when followed by a verb' symbolized \(p_{\text{verb}}\) and the constraint 'probability of contracting *is* when followed by a word that is not a verb' symbolized \(p\sim_{\text{verb}}\). Now, unlike the additive model, in the multiplicative models there are initially no mathematical conditions placed on the relations among the constraints in a particular family, other than that each be a number between zero and unity.

As we have just mentioned, it is assumed in the multiplicative models that the events represented by the probabilities within a constraint family are mutually exclusive. It is further assumed that each pair of probabilities from different constraint families correspond to independent events. In probability theory, two events are defined as independent just in case the probability of their joint occurrence is equal to the product of their individual probabilities. The probability of winning a daily double, or any other parlay bet, exemplifies this rule for combining probabilities. If there are five equally fast horses in the first race, the probability of winning the first race is one fifth, and if there are four equally fast horses in the second race, the probability of winning the second race is one fourth. To win the daily double, one must pick the winner in both of the first two races. The probability of doing this is the product of one fifth and one fourth, or one twentieth. The usual empirical interpretation of the mathematical definition of independence is that there is no causal connection between the two events. For example, if a racetrack is honest, the outcome of the second race is not affected by the outcome of the first, and conversely. To return to our example, it seems reasonable, at least as an initial assumption, to suppose that the influence of a

---

\(^7\) Actually the characterization 'mutually exclusive' is a slight over-simplification, but not one that creates any problem in the present context. For full details see Cedergren & Sankoff (1974: 340–1).

\(^8\) In general there may be more than a single constraint family at a given linear position, as would be the case in this example if we had also considered, say, the grammatical environment preceding the variably contracted *is*; this in fact is also an important conditioning factor (Labov 1969: 730 ff.).
preceding vowel on the contraction of is is independent of the influence of a following verb on the contraction of that same is, and conversely.

This model seems appropriate for variable rules. If there are \( n \) constraint families every particular token of application or non-application of the rule will involve one constraint from each family, for example: preceding vowel, following verb. We assume events of this kind are independent, which we agree with Cedergren and Sankoff is at least a profitable initial assumption. We may therefore think of the operation of a variable rule in producing a particular token in terms of a two-stage lottery of the following sort. We have \( n + 1 \) boxes, each containing a number of coins; the coins are not necessarily fair. The first box contains a single coin and corresponds to the 'input' constraint \( p_0 \). Each of the remaining boxes corresponds to a constraint family and the coins in it correspond to the individual constraints in that family. Each coin bears the name of the constraint to which it corresponds, for example, 'preceding vowel', and in addition has the words 'APPLY' stamped on one side and 'DON'T APPLY' stamped on the other. In the first stage, \( n + 1 \) coins are selected, one from each box corresponding to the conditioning feature present in the environment. In the second stage, all the coins are tossed, and the decision to apply or not to apply the rule is made on the basis of some preset criterion regarding the pattern of outcomes of the individual tosses. Cedergren & Sankoff (1974) have singled out two of the many possible decision criteria for study: (a) apply the rule just as if all the coins come up APPLY; (b) apply the rule if any coin comes up APPLY. Cedergren & Sankoff (1974: 337) have named these two decision criteria the 'multiplicative application probabilities' model and the 'multiplicative non-application probabilities' model respectively. We will abbreviate these labels 'applications' model and 'non-applications' model.

In the applications model we have the familiar parlay situation. The probability of the joint occurrence of a number of independent events is the product of their individual probabilities:

\[
(16) \text{ applications model } \ p = p_0 \times p_a \times \ldots \times p_n
\]

In the non-applications model we have a sort of negative parlay. The probability that one or more of the coins comes up APPLY is unity minus the probability that all coins come up DON'T APPLY. The probability that all coins come up DON'T APPLY is of course the product of the individual DON'T APPLY probabilities. Each DON'T APPLY probability is unity minus the corresponding APPLY probability. Hence, this model is given by

\[
(17) \text{ non-applications model } \ p = 1 - (1 - p_0)(1 - p_a) \ldots (1 - p_n)
\]

\[\text{[9]}\] A convenient way to distinguish among speakers or speaker types is to assign to each a distinct input constraint \( p_0 \), but we will confine our attention for the moment to the single speaker (or single homogeneous group) case.
ON THE LOGIC OF VARIABLE RULES

The factors of the second term of the right-hand side of (17) are the individual DON'T APPLY probabilities, while the factors on the right-hand side of (16) are the APPLY probabilities, which explains the choice of names.

Let us now consider how the underlying parameters of the multiplicative models are related to the probabilities of application generated by these models. That is, we wish now to perform some sample calculations illustrating the inner workings of the multiplicative models in the way that equations (6–8) above illustrated the additive model. We recall that in the additive model, we had to adopt some convention limiting the degrees of freedom in each constraint family. In particular, we assumed that the sum of the constraints in a family was always zero. We saw that when there are just two constraints per family, which is usual but not required in actual applications of the additive model, the result of this convention is that the two constraints are of the same absolute numerical value and of opposite sign. Thus, in our example, two of the five constraints were fixed by this assumption once we knew the other three, reducing the degrees of freedom from five to three (see equations (7)a, b), which made it possible for us to solve equations (6) by rewriting them as (8).

The defining equations in the multiplicative model are, as in the additive model, essentially summaries of a set of simultaneous equations, there being in both cases as many simultaneous equations as there are cells in the data table. So, in the case of the two-by-two data table, we have again four equations in five unknowns. (This result applies generally to any size data table, not merely the two-by-two case, a fact we will not prove here.)

What are the comparable assumptions made in the multiplicative models that yield a unique value of $P$? Cedergren and Sankoff assume that in the applications model $P_0$ is the probability that the rule will apply in the most favorable environment (i.e. the environment with the largest entry in the data table), and in the non-applications model they assume that $P_0$ is the probability that the rule will apply in the least favorable environment (Cedergren & Sankoff 1974: 341). No

[16] In the case of the applications model, applied to a two-by-two data matrix, the four equations, determined (or summarized) by the defining equation (16) are:

(i) $p = p_0 	imes p_{a1} 	imes p_{b1}$, e.g., $p = p_0 	imes p_v 	imes p_{verb}$
(ii) $p = p_0 	imes p_{a1} 	imes p_{b2}$, e.g., $p = p_0 	imes p_v x p_{nonverb}$
(iii) $p = p_0 	imes p_{a2} 	imes p_{b1}$, e.g., $p = p_0 	imes p_u 	imes p_{verb}$
(iv) $p = p_0 	imes p_{a2} 	imes p_{b2}$, e.g., $p = p_0 p_u x p_{nonverb}$

In the case of the non-applications model applied to a two-by-two data matrix, the four equations determined (or summarized) by the defining equation (17) are

(v) $p = 1 - (1 - p_0)(1 - p_{a1})(1 - p_{b1})$
(vi) $p = 1 - (1 - p_0)(1 - p_{a1})(1 - p_{b2})$
(vii) $p = 1 - (1 - p_0)(1 - p_{a2})(1 - p_{b1})$
(viii) $p = 1 - (1 - p_0)(1 - p_{a2})(1 - p_{b2})$
argument is given in favor of these assumptions, which have the consequence that in the applications model each of the variable constraints operating in the most favorable environment is equal to unity and in the non-application model that each of the variable constraints operating in the least favorable environment is equal to zero. Let us consider the applications model first. If the probability of application \( p \) for the most favorable environment is assumed equal to the input probability \( p_0 \), then the instantiation of the general rule (16) in this (most favorable) environment becomes

\[
(18) \quad p = p_a \times p_b \times \ldots \times p_n
\]

\[
1 = p_a \times p_b \times \ldots \times p_n
\]

So it must be the case that \( p_a = p_b = \ldots = p_a = 1 \). Hence, it is assumed that in the most favorable environment all the linguistic constraints are equal to unity and the input constraint alone accounts for all the variability. No support is offered by Cedergren and Sankoff for this seemingly arbitrary decision.

In the non-applications model, the assumption that the application probability \( p \) is equal to the input probability in the least favorable environment leads to the general rule (17) having the instantiation in this environment

\[
(19) \quad p = 1 - (1 - p)(1 - p_a)(1 - p_b) \ldots (1 - p_n)
\]

\[
\frac{p}{1 - p} = \frac{1}{1 - p} - (1 - p_a)(1 - p_b)\ldots (1 - p_n)
\]

\[
\frac{1}{1 - p} - \frac{p}{1 - p} = (1 - p_a)(1 - p_b)\ldots (1 - p_n)
\]

\[
\frac{1 - p}{1 - p} = (1 - p_a)(1 - p_b)\ldots (1 - p_n)
\]

Each factor on the right must of course equal unity and so each probability \( p_a, p_b, \ldots, p_n \) must equal zero. We are therefore assuming in this model that in the least favorable environment none of the linguistic factors makes any contribution to the probability that the rule will apply, the entire contribution being made by the input factor. Again, no evidence or argument is given by Cedergren and Sankoff for this assumption.

To return to our example of the two-by-two data table, the assumption has the desired effect of reducing the number of independent parameters to be estimated from five to three, since fixing \( p_0 \) equal to the observed \( p \) in the most (least) favorable environment has, as we have just observed, the effect of also determining that the variable constraints operating in that environment will both be fixed at unity (zero). But the cost of this move is to introduce an assumption with clear
empirical consequences but no empirical basis. In sum, our first criticism of the multiplicative models is that they may only be made to work by the introduction of an arbitrary assumption.

A second criticism of the multiplicative models involves a fact that the reader may have noted. In the applications model, every linguistic environment \( i \) has either no effect on the application of the rule \( (p_i = 1) \) or an enhancing effect \( (p_i < 1) \). Similarly, in the non-applications model, each linguistic environment \( i \) has either no effect on the application of the rule \( (p_i = 1) \) or an enhancing effect \( (p_i < 1) \). Thus an intuitively satisfying property of the additive model is lost: that some linguistic environments may have an enhancing effect on the probability the rule will be applied while others have an inhibitory effect. The additive model had, of course, to be abandoned for different and compelling reasons, as we have noted, but its replacement with the multiplicative models seems to introduce new problems. In the multiplicative models, the linguist is committed either to the assumption that every environment has an enhancing (or zero) effect (non-applications model) or to the assumption that every environment has an inhibitory (or zero) effect (applications model). These assumptions are tacit, but nevertheless unavoidable in the models proposed by Cedergren and Sankoff. No evidence or argument is offered in support of them. In addition we find them counter-intuitive, but whether or not the reader's intuition agrees with ours on this point, the best that can be said for these assumptions is that they are arbitrary.

Having pointed out certain arbitrary features of each of the variable rule models (additive, multiplicative applications, multiplicative non-applications), we consider now the methodology concerning which rule is to be applied to which set of data. The rule for choosing which variable rule model to use on which set of data is this: try all models and retain the one that fits the data best. This methodological principle is implicit throughout Cedergren and Sankoff's article and almost made explicit in several places. In connection with their analysis of R-spirantization in Panamanian Spanish they say, 'It remains to estimate the feature effects in the variable rule which yield a best fit to this data. Our procedure gives the values in Table 3 for the non-application probabilities [multiplicative non-applications model], which is the most consistent with these data' (Cedergren & Sankoff 1974: 345). Again, in connection with their analysis of que-deletion in Montréal French, 'In this case a multiplicative application model was used; this was far more consistent with the data than a non-application model, as measured by a chi-square comparison of predicted vs. observed frequencies' (Cedergren & Sankoff 1974: 348).

The method is to try all known models on each set of data and select the model that fits the given data best. But if the application of these different variable rule models to different corpora of data is to have linguistic significance, there must be some linguistic principle(s) according to which a given model is chosen for a given corpus, for example, 'Use the applications model in phonology and the
non-applications model in syntax', 'Use the additive model in isolating languages and the multiplicative models in agglutinative languages', or some such. Unfortunately, Cedergren and Sankoff offer no principled way of matching the type of model used to the type of data analyzed. Moreover, the principle of simply selecting the model that fits the data best raises the spectre of an ever-expanding number of models that might be tried (beyond the three specifically proposed to date) on a given corpus until one was finally found that fit well. For example, many other forms of the multiplicative model may be generated at will beside the applications and non-applications models. Remember that in the applications model, all the coins must come up APPLY and in the non-applications model at least one coin must come up APPLY. But given the fact that the methodological rule is to choose one’s model simply by what fits the data best – rather than according to some principle – there is nothing to prevent us from trying any number of multiplicative models: 'Apply if at least two coins come up APPLY', 'Apply if at least fifty percent of the coins come up APPLY', and so on. And we need not even restrict our choice to multiplicative models. If our choice of mathematical model is simply whatever fits the data best in each separate corpus, we might try exponential functions, power functions, indeed any function of the general form

\[(20) \ p = f(p_0, p_a, p_b, \ldots, p_n)\]

for a particular set of data until we find a good fit. The wider the range of models we allow, the more certain we are that some model in the family will fit an arbitrarily chosen corpus of data on the basis of chance alone. The principle for choosing among the three extant models is not only arbitrary in itself, but it renders the limitation to these three models arbitrary. That is, according to the Cedergren and Sankoff principle for matching model to corpus, there is no reason for the analyst to limit himself to the particular three models so far proposed. Rather this principle of choice is an open invitation to the analyst to devise more and more mathematical models until he has an armamentarium sufficient that one of them will model closely any set of randomly generated data. So we find arbitrariness, that is, lack of empirical motivation, not only in each of the variable rule models so far proposed but also in the principle determining which model is to be used on which corpus of data. Even if each of the models were individually immune to criticism, this equivocation would constitute a serious limitation on the realism of the variable rule method.

Rousseau and Sankoff have recently introduced a fourth variable rule model – the logistic model – which they now advocate using in all cases (Rousseau & Sankoff 1976; D. Sankoff 1976). Using a single model in all cases of course eliminates the arbitrariness of choosing whichever model fits the data best. The applications model tends to fit data best when the observed frequencies are all high (near unity), the non-applications model when the observed frequencies are
ON THE LOGIC OF VARIABLE RULES

low (near zero), and the additive model when they hover near 0.5. The logistic model tends to produce values of \( p \) close to those produced by the better-fitting one of the earlier models under each of the conditions just described. The logistic model has certain other desirable statistical properties whose discussion would be beyond the scope of this paper. Since at the time of this writing there does not yet exist a substantial corpus of empirical investigations employing the logistic model, it is not possible here to evaluate its utility or assess the degree to which exclusive use of the logistic model will answer the various questions raised regarding the variable rule method. The formula for the logistic model is

\[
(21) \quad p = \frac{p_0 \times \ldots \times p_n}{[p_0 \times \ldots \times p_n] + [(1 - p_0) \times \ldots \times (1 - p_n)]}
\]

In terms of coins and boxes, the logistic model has the following interpretation. Recalling that each box of coins represents a constraint family, one first draws from each box the coin corresponding to the particular constraint applicable in the environment in question as in the multiplicative models. One then tosses all the coins. If all coins come up APPLY the rule is applied; if all coins come up DON'T APPLY the rule is not applied; if neither of these two results occurs the coins are tossed again; and this is repeated until either all coins fall APPLY or all coins fall DON'T APPLY. An unsystematic survey of a few of the writers' acquaintances suggests that peoples' intuitions differ radically regarding the plausibility of this stochastic process as a likely model of psychological function. Before a hasty conclusion is drawn on this score, however, it is to be borne in mind that intuitions regarding plausibility are not the most reliable criteria for scientific judgment.

In the logistic model, no less than in the additive and multiplicative models, it is necessary to make certain (empirically unmotivated) mathematical assumptions to determine a unique value of \( p \). In this model one may either make the same assumptions as in the additive model or require that the mean of \( \log (p_i / 1 - p_i) \) within a constraint family be zero (Rousseau & Sankoff 1976).

4. VARIABLE RULES AND COMMUNITY GRAMMAR

If a variable rule is a rule of community grammar, then the mathematical assumptions inherent in published variable rule analyses are incompatible with a pattern of language change which has in fact frequently been observed. An assumption in each of the variable rule models so far proposed is that linguistic constraints and social constraints operate independently, that is, that there is no

\[\text{[11]}\] This section represents in abbreviated form an argument presented more fully in Kay (1978).

175
interaction between linguistic and social constraints. In fact, a very common pattern of observed language change, probably the characteristic pattern, involves precisely such interaction.

An attractive aspect of variable rule theory is the manner in which the linguistic constraints in the model represent propensities shared throughout a speech community while the social constraints and/or the input probabilities are what distinguish among social groups or individual speakers. On this interpretation, the effect of each linguistic environment on the probability of rule application is constant across speakers. The differences in application frequencies in a given linguistic environment observed in different speakers are attributed to social constraints and to the effect of speaker-linked input probabilities.

The variable-rule-as-community-grammar assumption has been propounded not so much in clear theoretical statements as in the large number of empirical studies that adopt it tacitly. In most variable rule studies (e.g. those reported in Cedergren & Sankoff (1974); G. Sankoff (1974); Wolfram (1974); Cedergren (1973); G. Sankoff & Cedergren (1972); Fasold (1972)) one simply applies one or several variable rule models to a data table representing the breakdown by linguistic and social environment of the application frequencies observed in the entire sample under study. In these studies no attempt is usually made to investigate whether or not linguistic constraints are shared; the method of analysis tacitly assumes they are. It is implied by the method that linguistic constraints are in fact shared throughout the speech community. The notion that linguistic constraints are shared community-wide and that speakers differ only with respect to their input probabilities is a plausible and attractive notion because it is the theoretical pillar of the doctrine that a variable rule expresses at once what is shared (the linguistic constraints) and what is not shared (the group or speaker constraints) in a community grammar.

Unfortunately, this notion of community grammar is shown in many of Labov's data to be contradicted by observed synchronic facts. We will demonstrate below

---

[12] There have been two exceptions of which I am aware. Cedergren & Sankoff (1974: 347) in their study of r-spirantization in 79 speakers of Panamanian Spanish, checked for uniformity of linguistic constraints by the following procedure. Using social-class-linked input probabilities derived from the variable rule analysis of the full sample of 79 speakers, they calculated expected frequencies for each environment for each speaker. Using a Chi-square test of goodness of fit, which they characterize as too stringent, they found a discrepancy between these predicted and the observed frequencies significant at the 5% level in about 20% of the informants.

Guy's recent study (1975) is a stronger exception in that Guy takes as a central issue the investigation of the degree to which linguistic constraints are shared across speakers His conclusion is that for the variable studied, deletion of final -d and -t in Philadelphia English, the relative strengths of linguistic constraints within a family are widely shared among speakers. (He does not attempt to assess the relative importance of different constraint families across speakers.) It is worth noting that the feature studied by Guy is not one that appears to be the focus of a change in progress.
ON THE LOGIC OF VARIABLE RULES

that it is also contradicted by a pattern of language change commonly observed in real communities.

Labov has often remarked that linguistic constraints are not in fact always uniform throughout a speech community. For example, in discussing the deletion of final dental stops among black teenagers, Labov indicates that the two chief linguistic constraints inhibiting deletion are (1) a morpheme boundary preceding the stop to be deleted (in effect making, the -t or -d the sole phonetic sign of the past-tense morpheme) and (2) a vowel beginning the following word. Labov points out that for different black teenage groups living only a few blocks apart in New York, the relative importance of these constraints differs. For example, in group session recordings when the first constraint family favors deletion and the second disfavors it, the Oscar Brothers group deletes about 90% of the time and the T-bird group only about 35%, whereas when the second constraint family favors deletion and the first disfavors it the T-Birds delete -t or -d about 90% of the time while the Oscar Brothers delete only about 55%. Thus for the Oscar Brothers the grammatical constraint is much more important than the phonological one while for the T-birds the reverse is true. Labov also plots average data on -t/-d deletion for several black and white adult speakers from various regions in the USA and England. In general, the same two constraint families promote/inhibit final dental stop deletion for all groups of English speakers with white and adult groups showing roughly the same range of variation of influence of the following consonant as black teenage groups but showing much more conservatism with respect to -t/-d deletion that would eliminate the past-tense morpheme (Labov 1972a: 44-8; 1972b: 78-85). Thus, in the case of -t/-d deletion, the linguistic constraints on the rule interact with the social constraints of group membership.

Having briefly considered some synchronic empirical limitations on the variable-rule-as-community-grammar theory, we take up some diachronic considerations. The uniform constraints assumption together with the independence of constraints property of all variable rule models jointly entail two empirical predictions with respect to situations of language change that are contradicted by at least most and perhaps all extant empirical data regarding cases of observed change in progress. These are (1) that the probability of application of the rule cannot increase (decrease) in a single environment to the exclusion of all others and (2) that if the relative order of constraint strengths is reversed for a single speaker, it must be reversed for all speakers. After demonstrating that these are in fact consequences of the theory, we will discuss a few empirical examples where they are violated.

In order to show that the two predictions just mentioned are in fact formal consequences of the variable rule theory, it is convenient to show first that, in an important sense, all variable rule models are variants, more precisely, order-preserving transformations, of the additive model. It will then suffice to show
that these predictions are entailed by the additive model, since the generalization
to all variable rule models will follow immediately from the fact that the other
models are monotone transformations of the additive.

For convenience, equations (4, 16, 17, 21), which define respectively the
additive, multiplicative applications, multiplicative non-applications, and logistic
models, are repeated as equations (22, 23, 24, 25):

\begin{align*}
(22) \quad p &= p_0 + p_a + \ldots + p_n \\
(23) \quad p &= p_0 \times p_a \times \ldots \times p_n \\
(24) \quad p &= 1 - (1 - p_0) \times (1 - p_a) \times \ldots \times (1 - p_n) \\
(25) \quad p &= \frac{p_0 \times \ldots \times p_n}{[p_0 \times \ldots \times p_n] + [(1 - p_0) \times \ldots \times (1 - p_n)]}
\end{align*}

where $p$ designates the probability of application, $p_0$ designates the input proba-
bility (which may vary by social group, including groups of one), $p_a$ designates
the constraint selected from the first linguistic constraint family, $\ldots$, and $p_n$
designates the constraint selected from the $n$th linguistic constraint family. As
has been pointed out independently by J. Kruskal (personal communication) and
D. Sankoff (personal communication), under the apparent diversity of these
models lies an interesting mathematical uniformity, namely that in each model we
can assign weights $k_0, \ldots, k_n$ to the linguistic constraints that correspond to the
probabilities $p_0, \ldots, p_n$ in such a way that the application probability $p$ is a
monotone function of the sum of the weights. That is for each model (22-5)
there exists a function $f$ such that

\begin{equation}
(26) \quad p = f(k_0 + \ldots + k_n)
\end{equation}

Consider first the additive model. Here we just take the weights $k_i$ as identical
to the $p_i$.

In the applications model (23) we let $k_i = \log p_i$. By taking the log of equation
(23) we get

\begin{align*}
(27) \quad \log p &= \log (p_0 \times \ldots \times p_n) \\
&= \log p_0 + \ldots + \log p_n \\
&= k_0 + \ldots + k_n
\end{align*}

Hence

\begin{equation}
(28) \quad p = \log^{-1} (k_0 + \ldots + k_n)
\end{equation}

In the non-applications model we set $k_i = \log(1 - p_i)$. Equation (24) may be
rewritten

\begin{equation}
(29) \quad 1 - p = (1 - p_0) \times \ldots \times (1 - p_n)
\end{equation}
ON THE LOGIC OF VARIABLE RULES

Taking the log of (29), we have

$$\log(1 - p) = \log[(1 - p_0) \times \ldots \times (1 - p_n)]$$
$$= \log (1 - p_0) + \ldots + \log (1 - p_n)$$
$$= k_0 + \ldots + k_n$$

Hence

$$1 - p = \log^{-1} (k_0 + \ldots + k_n)$$

and

$$p = 1 - \log^{-1} (k_0 + \ldots + k_n)$$

In the logistic model, if we take the weight \( k_i \) to be \( \log[p_i/(1 - p_i)] \), we obtain by reasoning analogous to the above

$$p = 1 - \frac{1}{1 + \log^{-1}(k_0 + \ldots + k_n)}$$

Thus we see that the four particular variable rule models so far proposed belong to a general class of models defined by equation (26) in which the application probability is always a monotone function of the sum of the weights assigned to relevant linguistic and social constraints.

The significance of equation (26) for the present discussion is that it allows us to see that the independence of constraints property is the same in all four models. Since all models are transformations of the additive model, all have the independence of constraints property, as defined in the additive model. Because of this property, any such transformed-additive model is incapable of accounting for a very common situation of linguistic change. Change often begins with an increase or decrease in frequency of rule application in a single environment, that is under a single combination of constraints, while the frequencies in other environments remain stable. Consider the case of two constraint families each containing two constraints: \((k_{01}, k_{02})\) the speakers, say, and \((k_{a1}, k_{a2})\) the linguistic constraints. Is it possible for the rule probability \( p \) to change in only one environment? For example, can \( k_{01} + k_{a1} \) change over time while the sums of weights for the other three environments \((k_{02} + k_{a1}, k_{01} + k_{a2}, k_{01} + k_{a2})\) remain constant? The answer is no, since if \( k_{01} + k_{a1} \) changes then either \( k_{01} \) or \( k_{a1} \) (or both) must change, and the change in one cannot be merely equal and of opposite sign to the change in the other. If, meanwhile, \( k_{02} + k_{a1} \) and \( k_{01} + k_{a2} \) are each to remain unchanged, then \( k_{02} \) must change to compensate for any change in \( k_{a1} \), and \( k_{a2} \) must change to compensate for any change in \( k_{01} \). But then \( k_{02} + k_{a1} \) must change by an amount equal and of opposite sign to the change in \( k_{01} + k_{a2} \). This shows that we cannot have a change in one environment only.

This effect is independent of the variable rule model chosen and, of course, of the procedure used to estimate the constraint values from the data. As long as the
number of different environments is large enough in comparison with the number of constraints – and no model is of any interest otherwise – a change in rule probability in one environment implies a change in at least one other environment. According to the mathematics of all the variable rule models, an increase or decrease in application probability in a single environment, to the exclusion of all others, is an impossibility. This theoretical impossibility is, however, a frequent concomitant of actual linguistic change.

A second, and more obvious, consequence of the independence of constraints property is that if constraints become reordered for any speaker, they must be similarly reordered for all speakers. This theoretical constraint is also violated by data of linguistic change in progress.

We have shown that none of the proposed variable rule models accounts for interaction of speaker and linguistic constraints. In particular, we saw that (a) the theory implies that there cannot be a shift of a single individual’s or group’s usage in a single linguistic environment while other groups and environments remain stable, and (b) the models imply that if the relative strengths of two linguistic constraints become reordered over time for one speaker or group of speakers, they must be similarly reordered for all groups of speakers in the community to which the variable rule applies. We now consider some data.

Perhaps the first of the modern studies of variation and change in progress that employed actual counts of linguistic tokens to get frequencies of rule application was Labov’s study of the centralization of the vocalic nucleus [a] of the diphthongs [aw] (bout) and [ay] (bite) in Martha’s Vineyard. Here it was found that the centralization of [a] experienced an initial jump in frequency in the environment [−y] among the fisherman of the old Yankee town of Chilmark. Thus proposition (a) is violated by the data. From here the change spread to the linguistic environment [−w] and to other speaker groups in the community, the Portuguese and Indian inhabitants, but not independently in the two dimensions: ‘In these two ethnic groups [Portuguese and Indian], centralization of (aw) overtook and surpassed centralization of (ay)’ (Labov 1972b: 525). Thus proposition (b) is also violated by these data.

If the raising of front and back peripheral tense vowels in New York is considered as a uniform linguistic process with [+ front] and [+ back] functioning as differential linguistic environments in a single variable rule, then there is interaction between these conditioning linguistic environments and the ethnic differentiation, Jewish versus Italian, with the former favoring the front and the latter the back vowels (Labov 1966: 11 and passim). Labov has more recently reported (1972d: 58ff and passim) that with respect to deletion of post-vocalic [r] younger upper-middle-class New Yorkers are dramatically adopting the non-deleted pronunciation in relaxed style without any significant comparable shift in other social classes. The former example constitutes a violation of proposition (b) and the latter of proposition (a).
ON THE LOGIC OF VARIABLE RULES

Several more detailed examples are discussed in Kay (1978), where the argument is made that observations of actual linguistic change in progress show it characteristically to involve the interaction of linguistic and social constraints. As we have seen, all variable rule models ordain independence (non-interaction) of all constraints and the variable-rule-as-community-grammar theory requires linguistic constraints to be shared throughout the speech community and thus specifies independence of social and linguistic constraints. In short, observed language change frequently, perhaps invariably, involves interaction of linguistic and social constraints while the variable-rule-as-community-grammar model prohibits this interaction. Perhaps this fact has received less than complete recognition because it was the same person, Labov, who (a) invented variable rules (1969), proposed the variable-rule-as-community-grammar model (1972a), (b) pointed out that variable constraints are not always shared throughout the community (1972a) and (c) documented in the finest detail that pattern of interaction between linguistic and social constraints that vitiates the variable-rule-as-community-grammar model (1972a, b, c and elsewhere).

5. SUMMARY AND CONCLUSIONS

In the introduction we stated two frequently encountered a priori arguments against the variable rule methodology and attempted to refute them. The first rejects the variable rule methodology – and by implication any study of comparable data – on the grounds that variable rules govern token frequencies while generative grammar does not countenance token frequencies. We agree that variable rules are not generative rules of a new sort but an entirely different kind of logical object and that generative grammar indeed does not countenance token frequencies. But we are convinced by empirical work conducted within the variable rule paradigm that token frequencies often display clear patterns and that moreover some knowledge of these patterns forms part of the linguistic abilities of speakers. We conclude that, whatever the drawbacks of the variable rule formalism, studies employing variable rules have shown regularities in linguistic behavior that point to a serious lack in the generative paradigm, narrowly defined.

A second line of argument against variable rules which we rejected was based on assumptions about human psychology and about probability theory. The assumption that human beings cannot assess probabilities and behave in accord with them in a natural and unconscious manner appears to be supported by no empirical evidence and does not seem to us plausible a priori. Moreover, experimental evidence to the contrary exists. The argument to probability theory was that a speaker would have to have an internal counting device to keep track of the relative frequencies of linguistic variants that he had heard from his own or other lips in order to behave in accordance with variable rules. But this would require a kind of probability theory that would differ in remarkable and unspecified ways.

181
from ordinary probability theory, since the paradigmatic empirical examples of the familiar theory, such as coins, dice, decks of cards, and so on, are not possessed of memories.

In section 2, we introduced the additive model for variable rules. We found that this model requires strong numerical constraints on the underlying parameters, which Labov referred to as the principle of geometric ordering. These are required to keep the 'probability of application' numbers between zero and unity but are not motivated empirically. Cedergren and Sankoff's suggestion that one 'truncate' by transforming the probabilities less than zero to zero and those greater than unity to unity seems less a solution to this problem than a refusal to face it. Most importantly, as Cedergren and Sankoff were undoubtedly aware but unaccountably failed to point out, the numerical weights attached to linguistic constraints and speaker types in the additive model are not in any sense probabilities. The 'application probability' is of course an algebraic combination of these numbers, and so it is quite unclear in what sense if any it is accurate to call this number a 'probability'. Certainly the usual accoutrements of any normal probability model - a specified random experiment with a specified sample space and a probability measure defined on it - are entirely lacking in the additive model.

The multiplicative models, discussed in section 2, have the great advantage over the additive model of a real probabilistic foundation. They have, however, other drawbacks. The first is that to apply either of these models to actual data unmotivated empirical assumptions must be made. In the applications model, Cedergren and Sankoff assume that the observed frequency in the most favorable environment is equal to the input probability, that is, that the numerical value of each variable constraint operating in the most favorable environment is unity and that the input probability accounts for all of the variation in the rule's operation. No justification is given in support of this belief. It is not explained why the input probability accounts for all the variation rather than, say, half of it - or any other fraction. Similarly no argument is made why all the linguistic constraints operating in this environment are assumed to have equal effect.

In parallel fashion Cedergren and Sankoff make the assumption in the non-applications model that the linguistic constraints all have a numerical value of zero in the least favorable environment. This assumption is equally rich in undesirable consequences and equally devoid of empirical foundation. Both assumptions appear to be motivated only by the fact that without some additional assumptions like these, the underlying parameters cannot be computed from a data table.

A second drawback has to do with interpretation of the multiplicative models. In the applications model, every constraint, linguistic or social, has either no effect on the application probability or an inhibitory effect. In the non-applications model every constraint has either a zero or an enhancing effect. Thus in
neither of the multiplicative models can one depict situations in which some linguistic factors variably enhance and others variably inhibit the likelihood of application of the rule.

A third criticism of Cedergren and Sankoff's postulation of the two particular multiplicative models is that these two models are chosen arbitrarily from an indefinitely large family of models. We pointed out that the applications model amounts to tossing a coin for each variable feature and applying the rule only if all coins come up APPLY. In the non-applications model one applies the rule if at least one coin comes up APPLY. But there are indefinitely many possible models in this family: apply if at least half the coins come up APPLY, if at least two thirds come up APPLY, and so on. Taken together with Cedergren and Sankoff's procedure of using whatever model fits the data best, multiplicative applications, multiplicative non-applications, or additive, this raises the spectre of an indefinitely large number of multiplicative models, one of which can be found to fit any set of data well, and the choice amongst which appears to be without empirical foundation or theoretical interpretation. Even without considering the indefinitely large number of multiplicative models implied but not specified by the Cedergren and Sankoff procedure, the practice of applying whichever of the three models fits the data best introduces in itself a serious element of arbitrariness.

We mentioned the recent development by Rousseau and Sankoff of a fourth, logistic, model, noting that as yet this model has not resulted in enough published empirical applications for us to evaluate it fully. Certainly, if use of this model can be justified for all empirical applications to the exclusion of all others, the arbitrariness in choice of model in the previous methodology will be eliminated. This model, like the others, however, is based on empirically unmotivated numerical assumptions and its probabilistic interpretation will doubtless strike some workers as far-fetched from a psychological point of view. This in itself is not much of a scientific criticism, we repeat. In any case experimental or other empirical evidence that human brains can and do make instantaneously the required kinds of calculations — which in the coin model may require an indefinitely long repetition of tosses — has not so far as we know yet been adduced.

Finally, we considered problems arising from the interaction of the variable rule methodology and the notion of community grammar. If variable rules are to be employed as a technique to represent a presumed supra-individual, community grammar, it would appear that the uniform constraints assumption is unavoidable both theoretically and methodologically. That is, if one does not assume that linguistic constraints are shared by all speakers in the community: (1) the notion of community grammar seems to make no sense; (2) the actual study of variable rules in the community breaks down into an individual rule for each speaker. But we found that adoption of the uniform constraints assumption leads to two consequences which are violated by the extant empirical data on language.
change in progress: (a) a given speaker or group cannot change his usage in a given linguistic environment to the exclusion of other speakers and environments and (b) if one speaker or group reorders the relative strengths of two linguistic contrasts, all speakers must do likewise.

Having criticized the variable rule methodology, we will end nevertheless on a positive note. First, in criticizing certain tacit assumptions involved in their formalism, we do not intend to belittle the importance of the work of Labov, Cedergren, D. Sankoff, G. Sankoff, Fasold, Wolfram and many others not explicitly cited regarding the observed regularities in token frequencies for a variety of linguistic variables studied in a wide range of linguistic and social environments. Also the work of variationists such as Bailey, Bickerton, DeCamp and others who do not accept the full variable rule methodology are relevant here. All these workers have demonstrated that the token frequencies of variable linguistic features do in fact show patterning and have made a strong case that some control of this patterning and its social significance is part of the linguistic ability of members of speech communities. They have pointed out an important area of linguistic data that was previously ignored and have moreover shown that there is regularity to be observed and subjected to theoretical understanding in these data.

Whereas Labov (1969), Cedergren & Sankoff (1974), and G. Sankoff (1974) have attempted to stretch Chomsky's notion of linguistic competence (see for example Chomsky (1965: 3f), and of course many of his other writings) to cover these facts, it seems to us that the quantitative patterning of token frequencies discovered by these and other empirical workers on language variation are better viewed as showing limits on the usefulness of Chomsky's notion of linguistic competence in particular and, more generally, on the strictly generative approach to language. As a formal theory of language, generative grammar takes the primary job of the linguist to be the specification of a function that enumerates a set of sentences as types. It seems inherent to Chomsky's conception of language, and thus his notion of linguistic competence, that language be viewed as a set of abstract objects called sentences and conceived as types, each of which is thought of as an ordered couple that pairs a sound sequence with a meaning. The primary job of the linguist on this view is to specify as clearly as possible the nature of the function that generates this set of ordered pairs. The specification of this function is an account of the speaker's competence.13

[13] Of course many linguists who consider themselves part of the now quite broad tradition deriving from the work of Chomsky — those whose work is in some sense either generative or transformational or both — do not accept the narrow limitations of the subject set by Chomsky in his most formally oriented writings, for example, '... a really insightful formulation of linguistic theory will have to begin by a determination of the kinds of permitted grammatical rules, and an exact specification of their form and the manner in which they impose structural descriptions on each of an infinite set of grammatical sentences' (Chomsky 1961: 6). As Green has recently noted '... as anyone
ON THE LOGIC OF VARIABLE RULES

The notion of a linguistic token plays no part in this account. Token frequencies might for Chomsky be a part of the data of a theory of linguistic performance, but so far as we know he has said nothing in print on this point. In any case it seems impossible to alter Chomsky's view of linguistic competence to account for observed patterns in token frequencies without distorting the original concept beyond recognition and usefulness. This is not, of course, to say that data on patterns of token frequencies are not important and interesting data for the linguist, nor to say that the theoretical notion of linguistic competence is without value, but it is probably a mistake to try to extend the formal theory known as generative grammar to accommodate data on token frequencies. Generative grammar is a formalism specifically designed to account for the systematic nature of the set of linguistic types in a language, and we doubt that the theory can be patched by the addition of variable rules to cover the data on token frequencies. As we argued in the introduction, variable rules are not 'like optional rules only a little more specific' but are an entirely different kind of conceptual object. At present it would seem preferable to acknowledge that we have no formal theory of variable data on linguistic tokens than to attempt to graft an account of these data onto a formal theory that was specifically designed for other and contrary purposes. It appears that there simply does not exist currently any formal theory that comes reasonably near to giving a coherent account of the systematicity observed in respect to linguistic tokens.

We have argued that the variable rule methodology represents a premature rush to formal theory, but in criticizing certain of the logical properties and consequences of the particular formal theory we intend no criticism of the excellent empirical work that has prompted it. The quantitative study of variable linguistic features as they are distributed in space, social space and time and as they covary with each other remains, we feel, a major research strategy to follow in seeking a deeper understanding of language change and of the relation of language to the society in which it is spoken. It appears, however, that further empirical work needs to be done before the field will be ready for a fully formal theory of variability in token frequencies that is free of the drawbacks pointed out above.

Variable rule analysis may well serve as an important tool in this empirical work. Here variable rule analysis is not thought of as grammar construction but as the application of a statistical technique, much like traditional analysis of variance but better adapted to the kind of data that arises in empirical linguistic investigations in which many cells of the data table have few or no cases. Viewed thus, as a descriptive and analytic statistical device for the reduction of data on the distribution of linguistic tokens across linguistic environments, speakers,

who reads the journals and the CLS volumes must be aware, describing language use has become a concern of many linguists who are, in the broad sense, transformational grammarians' (1977: 407).

185
styles, etc., variable rule analysis may contribute importantly to our increased understanding of linguistic variation, its interaction with social factors, and its intimate involvement in linguistic change.14

REFERENCES


14 The issues raised by Kay and McDaniel are of such importance that a response was invited. It appears in the following paper by Sankoff and Labov, which Kay and McDaniel of course did not see. (D.H.)

186
ON THE LOGIC OF VARIABLE RULES


