On the Uses of Variable Rules
Author(s): David Sankoff and William Labov
Source: *Language in Society*, Vol. 8, No. 2 (Aug., 1979), pp. 189-222
Published by: Cambridge University Press
Stable URL: http://www.jstor.org/stable/4167071
Accessed: 15/12/2014 19:42

Lang. Soc. **8**, 189–222. Printed in Great Britain

On the uses of variable rules

DAVID SANKOFF AND WILLIAM LABOV

Centre de recherches mathématiques, Université de Montréal & Department of Linguistics, University of Pennsylvania

PREFACE

The introduction of variable rules ten years ago has provoked a variety of critical reactions; among these Kay & McDaniel's (1979) review appears as a clear and penetrating study of many issues neglected or unresolved in earlier discussions. It is refreshing to observe an approach to the problems of variation and sociolinguistics relatively free from the ideological constraints that other critiques have inherited from formal linguistics (e.g. Bickerton 1971, 1973; Gazdar 1976). They bring to the discussion of variable rules a clarity gained by several years' reflection on the early statistical approaches, together with a certain distance from current sociolinguistic methodological developments and problems. At the same time, there are some attendant disadvantages of such a distance; these appear in their treatment of the work that preceded probabilistic models, in their lack of attention to the interaction between the practical aspects of linguistic data analysis and the evolution of theoretical concerns, in their misunderstanding of certain mathematical facts, and in their neglect of the more recent developments over the past five years. The K & M analysis may best be evaluated as a reaction to the stage in variable rule analysis around 1971–4 when the first probabilistic models were being proposed and tested.

It may therefore be helpful to make some comments which place the K & M analysis in a longer and broader perspective, relating the issues they raise to current advances in variable rules. The first section will trace our original motivations for the development of variable rules, and the second will clear up some of the problems encountered by K & M in attempting to understand and interpret the statistical theory. This will involve a brief exposition of mathematical developments to date. The third section will deal with how variable rules can reflect the heterogeneity and homogeneity of the speech community. In the fourth section, we turn to the relation of variable rules to generative grammar, and associated theoretical concerns.

Language in Society is not the appropriate forum for the exposition of purely statistical matters. It seems clear, however, that the sociolinguistic analysis of linguistic variation is destined to require an increasing reliance on statistical methodology of various kinds (see e.g. papers in D. Sankoff 1978a). Thus we felt it worthwhile that the fundamental aspects be discussed in a mathematically correct way, using established statistical terminology and criteria. We will try to follow K & M's example in making our presentation as self-contained and as readable as possible, but in treating the issues they raise, we will frame them in terms of well-known concepts and procedures of statistics and data analysis.

1. A HISTORICAL SKETCH

K & M's discussion of the development of variable rules begins with the 1969 treatment of contraction and deletion of the copula. It may be useful here to underline the extent of the difference between this and previous sociolinguistic studies, and why a new mode of notation (variable rules) was introduced to bolster the earlier description of sociolinguistic variation in terms of categories (linguistic variables).

The work in South Harlem on word-final *-t,d* deletion and on contraction and deletion of the copula constituted the first quantitative examination of internal constraints on linguistic variation (Labov, Cohen, Robins & Lewis 1968). Previous studies dealt only with the social distribution of particular variants.[1] Basically, they considered the underlying grammar or phonology as given and examined variation in the output, though there were ample indications that current discrete representations of that underlying grammar would have to be modified as time went on. But the Black English Vernacular was a different matter: here one could recognize the output, and trace the surface variation, yet not know the shape of the underlying grammar, or what alternations in that grammar were responsible for the variation. The surface realizations of BEV were different enough to require some different organization of the rule system. It was not clear first how radical that difference was – this in fact was the main issue on the table.

Earlier statements about BEV and related Caribbean grammars were cast in categorical terms, and discussions of origins and relations to other dialects were based on qualitative arguments and anecdotal evidence.[2] The studies in South Harlem, on the other hand, used quantitative relations to establish the existence of a past-tense morpheme, the non-existence of a third-singular marker, the existence of an underlying copula and processes of contraction and deletion. Ordered rules turned out to be a valuable way of organizing the data, and helped show that two closely related processes of /l/ and /r/ vocalization were separated

[1] The study of Martha's Vineyard may be considered an exception here, since some semi-quantitative information on the ordering of internal constaints was given (Labov 1963, 1972a: 20). But this information was not systematically developed and played no role in the discussion of linguistic change.

[2] See Bailey (1969) and Stewart (1970). These qualitative statements may have overshot the mark as characterizations of BEV, but the Creole origins of BEV which they argued for have been further confirmed in later research.

by an intervening process of consonant cluster simplification.[3] Auxiliary contraction and deletion were found to be clearly linked to some form of stress assignment and vowel reduction as postulated in ***The sound pattern of English*** (Chomsky & Halle 1968). Though introspective arguments on auxiliary contraction have been advanced to complicate our view of this process (Zwicky 1970), the relation to stress and vowel reduction still seems valid.

The quantitative aspect of this work lay not only in the heuristic manipulation of linguistic production data as part of a discovery procedure but also in the recognition and proof of apparently distinct reduction processes as parallel phenomena representable by a single rule. Further, important generalizations of well-known types of categorical relationships between constraints and variables were found to be expressible only in quantitative terms.

The variable rule analysis was not then presented as an isolated borrowing from the generative format, but as sixteen intimately related rules of morphophonemic reduction, some categorical, others variable. There have been a number of advances and revisions in this set of rules (Fasold 1972; Wolfram 1974), and other studies of the BEV tense and aspect system show categories which are more remote from other English dialects (e.g. Rickford 1975); but the quantitative evidence on internal constraints on contraction and deletion rules has been confirmed by a large number of other studies (e.g. Mitchell-Kernan 1969; Wolfram 1969; Legum *et al.*, 1971; Baugh, to appear). The variable format has turned out to be a convenient and revealing way of representing underlying complex internal relations in the grammar, relations which are systematic and shared among speakers, though quantitative rather than categorical or discrete.

The formal conventions for interpreting these rules were not originally phrased in terms of probabilistic models or statistical theory and method. K & M present a view of the algorithm proposed at that time to account for multiple feature effects on rule frequencies, as if it were conceived then in terms of a probabilistic model, but no such element was present. On the basis of experience with many stable relations of more-and-less in the speech community, there was ample motivation to enter these relations into a rule format. But the fundamental distinction between frequencies and probabilities and other questions of model choice, statistical estimation theory, and hypothesis-testing had not yet assumed importance in that work. Thus K & M's damning of the additive model, for containing parameters which are supposedly uninterpretable since they are not probabilities, can best be characterized as unedifying hindsight. Mathematically, it is also wrong, since there is no logical reason nor even any statistical tradition

[3] The vocalization of /r/ and /l/ are parallel rules of English phonology. Yet at the point where *-t,d* deletion applies, /r/ has been vocalized but not /l/. Thus *cord* does not have a cluster as far as *-t,d* deletion is concerned, and behaves like other forms with a single final /d/; but *cold* does behave like other CVCC forms and the probability of deleting the final consonant shows that at this stage in the derivation, /l/ is a consonant. In BEV, as in many other dialects, this postvocalic /l/ is consistently vocalized in the final output.

for all the parameters of a probabilistic model to be themselves probabilities, though they may have, as is the case here, very clearcut interpretations. K & M's vehemence on this point extends to a critique of Cedergren & Sankoff (1974) for 'being undoubtedly aware but [having] unaccountably failed to point out' this supposed drawback of the additive model, though these latter authors do in fact explicitly refer to '...additive models, where the coefficients are not automatically interpretable as probabilities' (339).

In the early work, it was of course obvious that the tendency of the additive model to predict frequencies greater than 100% or less than 0% would have to be limited by truncation or some other device.[4] Some of the earlier data suggested that a geometric model of constraint relations did in fact hold for the data, as described in K & M. But it quickly became clear that such a geometric relation, with a stable hierarchy of internal constraints, could not apply generally. Even in the first analysis of *-t,d* deletion, it appeared that there were speakers who had the two major constraints evenly balanced. Wolfram, however, has continued to make effective use of the postulate of geometric ordering in a number of analyses and Fasold (1978) claims that it is a widely valid principle.

A more important limitation of the earlier work was the absence of any mode

[4] K & M's discussion of necessary and sufficient conditions for a rule probability p to fall, for any set of constraints, in the interval [0, 1], is mathematically confused and erroneous. Their conditions

(K & M 10) $0 \leq p_0 \leq 0.5$

(K & M 11) $p_0 \geq p_i$ for $p_i \, \varepsilon \{p_a, p_b, \ldots, p_n\}$

(K & M 12) Relabelling families so that $p_1 \geq \ldots \geq p_n$,

$$p_i \geq \sum_{i+1}^{n} p_j \; (i, j \, \varepsilon \{1, 2, \ldots, n\})$$

do not jointly express a sufficient condition as they claim ($p_0 = 0.4, p_1 = 0.3, p_2 = 0.2$ is a counterexample, since these values satisfy the three conditions but allow $p < 0$ in one context). Nor is their condition (K & M 12) a necessary condition, as they claim ($p_0 = 0.5, p_1 = 0.1, p_2 = 0.1, p_3 = 0.1$ is a counterexample since these value combine to give values of p only within [0, 1] but do not satisfy condition (K & M 12) as stated).

A correct formulation would be that a necessary (and sufficient) condition for p to fall in [0, 1] for all contexts is

$$(1.1) \qquad \sum_{i=1}^{n} p_i \leq p_0, \qquad \sum_{i=1}^{n} p_i \leq 1 - p_0.$$

A much more restrictive sufficient condition is given by replacing the first of the inequalities in (1.1) by

$$(1.2) \qquad \sum_{i+1}^{n} p_i \leq p_i, \quad i = 0, 1, \ldots, n-1$$

this being the only mathematically meaningful interpretation of Labov (1969: 740–1). The reader should also be alert to the fact that K & M's explanation of the meaning of their conditions (K & M 10), (K & M 11), and (K & M 12) confuses the idea of logical 'necessity' and 'sufficiency'.

of multivariate analysis, which could take into account the large number of intersecting constraints that apply simultaneously. Sometimes the effects of competing constraints could be controlled by subdividing categories to the point where the effect of one constraint could be examined separately while the rest were held fixed. But in many cases, the data ran out before reaching that point. Categories were then collapsed which afterwards proved to be quite distinct (Baugh, to appear), or percentages were shown across categories without taking into account the possibility of skewed distributions within those categories (Labov 1975).

The South Harlem work, then, served to prove the desirability and significance of incorporating quantitative considerations into grammatical description, to discover rule application frequencies as the fundamental data of variation study, to show how the quantitative effect of the linguistic environment on a rule can and should be decomposed into a combination of effects from the various constraints present in this environment, and to propose terminology and notation enabling variable rules to be easily integrated into the formalism of generative grammar. The statistical and probabilistic aspects we have alluded to were not really investigated before the analysis of Cedergren's data on Panamanian Spanish (Cedergren 1973a).

The sheer volume of the Panamanian data necessitated a systematic and computer-implemented analysis. Statisticians generally use analysis of variance (ANOVA) or multiple regression when trying to decompose a quantitative phenomenon into a number of cross-cutting effects. These methods consist of:

(i) a model for such decompositions, namely an additive, linear model,
(ii) a parameter estimation criterion, namely the minimization of the sum of squared differences between the observed quantities and their respective values predicted by the linear additive formula,
(iii) a rapid computational formula for finding the estimates satisfying this criterion – especially rapid in the case of ANOVA, and
(iv) a series of tests for the significance of parameter differences – especially sensitive and revealing in the case of ANOVA.

Because of the particular nature of linguistic data, both (i) and (ii) were felt to be inapplicable. Frequencies of rule application range only between 0% and 100%, but the usual additive model mentioned in (i) does not necessarily predict values respecting this constraint, formalized above as (1.1), and so other models which do must be tried. As for (ii), data on a linguistic variable collected from a speech sample tend to be very unevenly distributed among the various possible contexts, a situation in which the least-squares type of criterion loses its good statistical properties, becoming subject to considerable and unpredictable inaccuracies. Even when the sample can be gathered under controlled experimental conditions, gaps in the data usually prevent any even distribution, since many combinations of contextual features are not realized in the language. This forces us to rely on a more fundamental estimation criterion, that of maximum likelihood, which retains its statistically desirable properties despite poorly distributed data, but whose computation is much more difficult. Thus aspect (iii) of ANOVA and regression analysis is also lost, and with it the tests mentioned in (iv), these being byproducts of the computational procedure.

In the next section we discuss the various solutions adopted to provide an analysis similar to regression and ANOVA but particularly adapted to the peculiarities of linguistic variation data. It goes without saying that the methods in question do not apply to all linguistic data. On one hand, there are qualitative data which concern the existence or non-existence of certain forms which are particularly important in the study of little-known languages, the first steps in historical reconstruction, and the quest for linguistic universals. On the other hand, the study of vowel shifts in progress provides continuous quantitative data as the dependent variable. Here multiple regression and ANOVA are particularly appropriate, especially for the study of social influences on linguistic change (Labov *et al.* 1978). Our discussion here concerns the very large field of linguistic variation where the dependent variable is a discrete choice, usually binary, and usually formalized as a rule with variable application.

2. DATA ANALYSIS

Models

In what follows we will conserve as much as possible the notation of K & M except that it will be clearer to use β_a rather than p_a to denote the effect of parameter *a*. In addition, while we will use β_a to represent the effect of an unspecified member of family '*a*', we will also use $\beta_{a1}, \beta_{a2}, \ldots$ to represent the effects of particular constraints *a*1, *a*2,... in that family.

As K & M point out, the additive model

$$(2.1) \qquad p = \beta_0 + \beta_a + \ldots + \beta_n$$

for accounting for the effects of environmental features on the rule probability p has to be modified to ensure that the parameter values produced by the estimation procedure cannot combine, in any environment, to give a p greater than one or less than zero.

This can be done in many ways. For example, we may impose the necessary and sufficient condition (1.1) in footnote 4 on the β_i, with or without the strong geometric ordering principle (1.2). Alternatively, we could truncate, i.e. set

$$(2.2) \qquad \begin{aligned} p &= 0 && \text{if } \beta_0+\beta_a+\beta_b+\ldots+\beta_n<0 \\ p &= 1 && \text{if } \beta_0+\beta_a+\beta_b+\ldots+\beta_n>1 \\ p &= \beta_0+\beta_a+\beta_b+\ldots+\beta_n && \text{otherwise} \end{aligned}$$

Another solution is to make some function of p additive, instead of p. Cedergren & Sankoff (1974) suggested

$$\text{(2.3)} \qquad \log p = \beta_0 + \beta_a + \ldots + \beta_n$$

$$\text{(2.4)} \qquad -\log(1-p) = \beta_0 + \beta_a + \ldots + \beta_n$$

might be suitable functions. A further possibility, discussed by Rousseau & Sankoff (1978a), and which replaced the others in general use after 1974, is given by

$$\text{(2.5)} \qquad \log \frac{p}{1-p} = \beta_0 + \beta_a + \ldots + \beta_n$$

And there are many other possible functions which have been used by data analysts in other contexts, though not in linguistics (Cox 1970).

The idea behind the use of such functions is that even allowing the estimation procedure to estimate very large magnitudes for the β, this should be consistent with a value for p between zero and one.[5]

Constraints on the model

K & M state that additional constraints[6] are necessary on any such model because 'without some additional assumptions like these the underlying parameters cannot be computed from a data table'. They criticize the constraints chosen by Cedergren & Sankoff (1974) and Rousseau & Sankoff (1978a) as 'unmotivated empirical assumptions...' 'rich in undesirable consequences'. All of these remarks are mathematically incorrect. The constraints, of form

$$\text{(2.6)} \qquad \sum_i \beta_{ai} = \sum_j \beta_{bj} = \ldots = \sum_k \beta_{nk} = 0$$

for models (2.1) and (2.5), where $a_1, a_2, \ldots$ are the different possible environmental features or constraints in the 'a' family, etc.; or

$$\text{(2.7)} \qquad \beta_0 \leq \max_i \beta_{ai} = \max_j \beta_{bj} = \ldots = \max_k \beta_{nk} = 0$$

$$\text{(2.8)} \qquad \beta_0 \geq \min_i \beta_{ai} = \min_j \beta_{bj} = \ldots = \min_k \beta_{nk} = 0$$

for models (2.3) and (2.4), respectively, are not responsible for the computability of the parameters, which can be computed without them. They are not empirical

[5] Models (2.3), (2.4), and (2.5) have generally been presented in forms which make less transparent their relationship to each other and to (2.1). Defining another set of parameters $p_0, p_a, \ldots, p_b$ where $\log p_i = \beta_i$, equation (2.3) takes the form (K & M 16), the multiplicative applications model. Similarly, defining $-\log(1-p) = \beta_i$, equation (2.4) becomes (K & M 17) the multiplicative non-applications model. The transformation $\log(p/1-p) = \beta_i$ converts (2.5) into (K & M 21), the logistic model. Note that (K & M 33) is an erroneous representation of this model; a corrected version is

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

[6] K & M's constraints 7a and 7b are wrong and inconsistent with the rest of their discussion. Equation (K & M 7a) for example should be $p_{-V} = -p_{+V}$.

assumptions, nor do they have any substantive consequences whatsoever. The point is that in postulating any transformed-additive model such as (2.1)–(2.5) to account for a set of data, and in estimating parameter values, we cannot in any logical or mathematical way claim to be interested in the individual parameter values in a constituent family, for example, $\beta_{a1}, \beta_{a2}, \ldots$, but only in the differences[7] between these parameters: $\beta_{a1}-\beta_{a2}, \beta_{a1}-\beta_{an}, \beta_{a2}-\beta_{an}, \ldots$. This can be seen by taking any set of estimates for the parameters, adding an arbitrary number, say $126\frac{1}{2}$, to all the β_a, subtracting the same number from all the β_b, and leaving the rest of the parameters unaltered. Since

$$(2.9)\quad \beta_0+(\beta_a+126\tfrac{1}{2})+(\beta_b-126\tfrac{1}{2})+\ldots+\beta_n = \beta_0+\beta_a+\beta_b+\ldots+\beta_n$$

the altered set of estimates predicts the same probabilities as does the original set and is hence as good or as bad a set of estimates. This holds true in any of (2.1), (2.2), (2.3), (2.4), or (2.5). Of course there is nothing special about $126\frac{1}{2}$ and we could use other numbers to produce an infinite range of sets of estimates. What is important, however, is that in the altered set, the differences between the parameters in a family, for example, $(\beta_{a1}+126\frac{1}{2})-(\beta_{a2}+126\frac{1}{2}) = \beta_{a1}-\beta_{a2}$, are the same as the differences between corresponding parameters in the original set. These differences are uniquely determined by the conjuncture of model, data and estimating criterion and are the quantities of interest, however implicitly, in interpreting the results. The importance of the differences is perceived, though somewhat dimly, by K & M in their proposed 'quick test' for the additive model. The fact that, without constraints like (2.6)–(2.8), the values of individual parameters in a constraint family are free to wander together into the hundreds, thousands, or millions is no logical problem, nor does it impede computability of the estimates, but it is inconvenient notationally. Thus we impose (2.6), (2.7), or (2.8), which do not affect the parameter differences and are hence not empirical assumptions. They simply peg, arbitrarily, each constraint family around the same specific value (zero), leaving β_0 free to represent the overall average tendencies in the data. There is an additional notational convenience in imposing (2.6)–(2.8); that is, it permits the rapid comparison of effects *between* constraint families – this could be done anyway, by comparing intrafamily differences for different families, but with (2.6)–(2.8) the size of the smallest and largest parameters in a family gives a rough idea of the importance of that family in relation to other constraint families.

There remains the question of why (2.7) and (2.8) take the particular form they do. First of all, the models (2.3) and (2.4) do not in themselves guarantee that p remain between zero and one, nor would any constraint like (2.6) help, whereas (2.7) and (2.8) do. In practice, however, this guarantee would rarely be needed

[7] In terms of the transformed parameters p_i in the multiplicative models referred to in footnote 5 and (K & M 16), it is the ratio of parameters p_{a_1}/p_{a2} which is important. For (K & M 17) it is ratio $(1-p_{a1})/(1-p_{a2})$.

unless there were no data for certain important environments. A second reason is that (2.3) together with (2.7), or (2.4) together with (2.8), lend themselves to certain interpretations in terms of the probability analysis of independent events, on which K & M place particular stress, though they have been of little interest to variation theorists over the past five years. K & M criticize these assumptions[8] as being unjustified, unsupported, and arbitrary, and as having clear empirical consequences but no empirical basis. Indeed, as we have shown, they are mathematically arbitrary and have no empirical basis, but their consequences are not empirical but notational. Whether the assumptions are made or not has no substantive effect on the data analysis or linguistic consequences in terms of parameter differences. We have shown how to interpret these models as membert of an infinite class of transformed-additive models for which no probabilistic interpretation of the parameters is necessary. Thus K & M's claim is false; it is not true that the particular probabilistic assumptions they dwell on are unavoidable.

In discussing (2.7) and (2.8) we have strayed from the statistical concern of this section in order to demonstrate the irrelevance of K & M's critique, but we postpone further discussion of probabilistic interpretations until section 4.

The logistic-linear model, represented by (2.5), has several advantages over the previous three models and it is likely that variable rule analysis will continue on this basis, as it has for the last five years. For example, model (2.3) is most sensitive to the differentiation of strongly inhibiting constraints as a rule, while it tends to obscure favorable effects, i.e., where β_i is close to zero. Model (2.4) has a reverse property. The logistic model is evenly balanced in this respect.

Other desirable properties of this model become clear in the considerable literature and methodology based on it in the last few years in the field of statistics (Cox 1970; Lindsey 1975; Haberman 1974; Rousseau 1978). In linguistics this model has been adopted not only by the project on linguistic change and variation in Philadelphia (LCV) and the Montréal French group, but also by researchers in Washington, D.C. (Fasold 1978), Rio de Janeiro (Naro & Lemle 1976); Lemle & Naro 1977), Germany (Lüdicke 1977), and other centers.

Estimation

Having dealt with the statistical role of models, we turn to the problem of estimation. K & M do not discuss the advantages, over least-squares methods, of the procedures for obtaining maximum likelihood estimates. As we have stated, since the nature of linguistic structure is such that most cross-cutting sets of features must contain empty cells (and often the great majority of such cells are

[8] K & M derive (2.7) by assuming $\beta_0 = \log p$ in the most favorable environment, which they wrongly identify with the largest entry in the data table (this is not necessarily or even usually the case), and then using the assumption that all $\beta \leq 0$.

empty), the usual modes of multivariate analysis like ANOVA are not workable. Maximum likelihood methods are designed to operate on the actual number of occurrences rather than percentages for each speaker and to show an increased fit to the larger cells while giving proportionately decreased weight to small cells with one or two tokens. K & M's analysis of a fictional, highly unlikely, data set in terms of the solution of a set of simultaneous equations is misleading in this respect. Though they claim this is done for the purposes of exposition (presumably of the relationship of parameter estimates to data) and for exemplifying the internal mathematics of variable rules, it shows nothing about the nature of maximum likelihood estimates, or anything about the mathematics beyond what is already explicit in the model itself. The calculations may have served an autopedagogical purpose for the authors, but are irrelevant to real statistical estimation, which does not at all resemble their procedures.

Given *any* set of values for the parameters β_{a1}, β_{a2}, ..., β_{b1}, β_{b2}, ..., the likelihood of this set, as a function of particular data, is obtained by first calculating in each cell the value of p from the appropriate formula: (2.3), (2.4), or as is now prevalent, (2.5), and then, denoting by A the number of rule applications in that cell and T the total number of tokens, calculating

$$p^A\,(1-p)^{T-A} \qquad (2.10)$$

The product of all these terms of form (2.10) over all cells represents the likelihood of the parameter set. To find the set of parameter values which result in a maximum product for this likelihood, successive approximation methods of mathematical programming must be used. For the logistic model, the computer program VARBRUL 2 (Sankoff 1975) is in wide use. A more flexible version adapted to the PDP-11 system has been developed by the LCV project. The most sophisticated and powerful version, VARBRUL 3, has been implemented by Rousseau (1978).

Testing

One of the most important innovations of Rousseau's program, and one of the simplest (so that it is easily grafted on to other presently implemented versions), is the log-likelihood test for seeing if a constraint has a significantly different effect from another in its constraint family, or for seeing whether an entire constraint family should be retained or discarded from the model on the basis of whether it contains at least two significantly different effects. The maximum likelihood estimation produces a figure for the over-all likelihood of a given analysis at each iteration or successive approximation of the calculation. At the point where satisfactory convergence is reached, i.e. the estimates are sufficiently pinpointed, this likelihood can be printed out and compared with the value for another organization of environmental constraints, for example, with one constraint distinction less, or with one constraint family omitted from the model. Under the null hypothesis that the extra constraints have no effect, twice the difference in the logarithms of these likelihoods is distributed as a chi-square, with degrees of freedom equal to the differences in number of parameters estimated in one run as compared to the other.

Thus Weiner and Labov's analysis of the passive (1977) showed that preceding clauses with subjects coreferent to the underlying object of a sentence favored the application of the passive transformation as against an active sentence with a generalized pronoun. One such analysis compared all cases with a preceding coreferent subject against cases where the preceding subject was different. The resulting weights[9] in the factor group were 0.42 for the same subject cases and 0.58 for different subjects, with an overall log likelihood of −888.97. A second run distinguished sentences with a coreferent subject in only one preceding clause against those with strings of coreferent subjects in two or more clauses. The results were:

0.39 no coreferent subject preceding;
0.49 one clause with coreferent subject preceding;
0.62 two or more clauses with coreferent subject preceding
and a log likelihood of −885.54.

Twice the difference in these log likelihoods is 6.8, and a chi-square test with one degree of freedom shows that this difference is significant at the 0.1 level. Thus the additional precision in the description of the effect of a preceding string

TABLE I

Variable	Increase in likelihood achieved by adding one extralinguistic factor							
	To linguistic factors only				To linguistic factors plus other three extralinguistic factors			
	Socio-linguistic index	Education	Age	Sex	Socio-linguistic index	Education	Age	Sex
on/tu-vous	11.5	0.3	49.8	151.8	0.9	6.9	32.3	123.9
on/nous	58.4	37.3	43.5	1.0	15.3	4.8	84.4	0.4
on/ils	136.1	120.2	0	21.5	29.9	2.9	8.5	9.3

[9] For historical reasons and for the sake of comparability of different analyses, we generally follow the convention of not giving the β_i directly, but rather p_i, where

$$\log \frac{p_i}{1 - p_i} = \beta_i$$

The weights are thus restricted to the interval [0, 1].

can be said to add significantly to the determination of the choice of active vs. passive.

Laberge (1977) made extensive use of the log-likelihood test to evaluate the effects of various extralinguistic factors – age, sex, education and insertion in the linguistic market (D. Sankoff & Laberge 1978) – on three variables which involve the pronoun *on* as a variant in Montréal French. One of the variables, the *on/tu-vous* variable, has recently undergone a marked change whose chief locus is younger men. As seen in Table 1, both sex and age are statistically very significant factors, whether considered as the sole extralinguistic factors or whether combined with the others. The factors indicative of social level are barely significant and this only in an inconsistent way.

Turning to the second variable, *on/nous*, in which change has almost gone to completion, we see that a trace of this change remains in the highly significant age factor. Here, however, the sociolinguistic index shows a clearcut social stratification of the variable. The effect of educational level when used as the single extralinguistic factor is largely due to its correlation with the sociolinguistic index and perhaps its negative correlation with age, as it loses most of its significance when used together with the other factors. In contrast to *on/tu-vous*, there is no effect whatsoever of speaker's sex on the *on/nous* variable. The *on/ils* variable is the most stable of the three as can be seen from the relative insignificance of its age effect. There is clear social stratification as indicated by the significance of the sociolinguistic index and some degree of differentiation according to sex. Note that here we are assessing the importance of factors in terms of the statistical significance of their effects rather than the size (relative values of β_i or p_i) of their effects. Generally the two go together, but sometimes an effect appears larger or smaller than it should as an accident of statistically poorly distributed data, a problem which is controlled by examining significance levels.

Another capability of Rousseau's program, as used by Laberge, is the incorporation of continuous factors. Thus, rather than dividing the sample of speakers into age groups and a number of sociolinguistic or educational levels, with a separate factor assigned to each group and level, the actual age, years of schooling and sociolinguistic index can be incorporated directly into the model. Suppose z is a continuous variable such as speaker's age. Then (2.5) can be modified as follows

$$\log \frac{p}{1-p} = \beta_0 + \beta_a + \ldots + \beta_n + c(z-\bar{z}) \qquad (2.11)$$

where c is a single parameter and $\bar{z}$ is the average age. Thus for *on/nous* $c = 0.05$ per year in the direction of *on* whereas for the stable *on/ils* it is only 0.015 in the direction of *ils*. For the variable *on/tu-vous* which is currently undergoing change, it is only 0.016 per year in the direction of *tu-vous*, but as we will explain in section 3.2, this is an artifact of interaction between sex and age effects, with women moving in a direction opposite to men.

The log-likelihood test is based on certain approximations, but this does not limit its applicability to linguistic data sets with cells containing only one or a few tokens. This is an important improvement over earlier 'observed-versus-expected' chi-square tests. K & M's 'handy, though vague, quick check' is of no discernible value in these cases.

Choice of models

K & M's critique, though logically incorrect as concerns mathematical questions, brings up an important point concerning the disadvantage of postulating two alternative models for variable rules and choosing independently, for each data set, the one which fits best. However, K & M's discussion puts far too much emphasis on the selection of a 'best' model, which was in practice never a primary consideration. On the contrary, the main use of the various models was to locate stable and robust effects that appear in all models, and any data giving results that were highly model-dependent were considered insufficient for analysis. It is true that Cedergren and Sankoff did try to motivate the two models (2.3) and (2.4) in terms of probabilistic independence of constraints, rather than the broader type of data-analytic consideration which led to the adoption of (2.5), but these motivations have not proved to be instructive or to have any verifiable empirical consequence. It has been found that one particular model normally proves best for a given rule; model (2.4) consistently showed the best fit of observation and prediction for the aspiration and deletion of Spanish /s/, and for the contraction and deletion of the copula. Model (2.3) consistently showed the best fit for *-t,d* deletion. It is probable, however, that this sort of result is due to the distribution of one type of data towards 100% rule application and another towards 0%, rather than any more profound linguistic mechanism.

Returning to mathematical considerations, in their *reductio ad absurdum* exercise envisaging an arbitrarily large set of models (which is somewhat exaggerated, given that Cedergren and Sankoff only proposed two), K & M make a serious error. It is not scientifically unjustifiable, as they intimate, to start with an infinitely large class of models and use the data to single out the most appropriate. In fact, this is the basis for the MONANOVA procedure of Kruskal (1965) which actually selects from the infinite set of all continuously increasing functions of p, the one which leads to a best fit of a linear-additive model to a specific data set. This is done in a least-squares rather than a maximum-likelihood context, though an analogous procedure can be defined for linguistic data. The idea would be to compare the functions obtained for a series of data sets in the hope of finding some linguistically meaningful generalization.

For the present, however, the uniform use of the logistic model (2.5) seems the best strategy for systematic variation analysis suitable for comparison across different data sets and in different speech communities. The statistical and computational developments to date on variable rule models have brought us steadily closer to standard statistical approaches to the analysis of variation, still preserving the ability to deal with the special characteristics of linguistic data.

3. THE INDIVIDUAL AND THE SPEECH COMMUNITY

K & M's discussion of variable rules, generally quite lucid, becomes less than clear when they turn to the issues involving the individual and the community. They consider here jointly two 'assumptions': one described as the 'variable-rule-as-community-grammar assumption' and the other that 'linguistic constraints and social constraints operate independently'.

Both of these assumptions are linguistic interpretations of a mathematical assumption used in the estimation of constraint effects as evidenced by the data. As such, they are not correctly interpretable as linguistic hypotheses, tacit or otherwise, on the part of variable rule users. The assumption of the additivity of constraints may be correctly thought of as a statistical null hypothesis, which is methodologically necessary even when linguistically we are convinced there are non-additive interactions, and we want to prove and evaluate these from the data. Given the inapplicability of ANOVA and related methods to linguistic data, and hence the inaccessibility of their detailed recipes for detecting and unraveling constraint effects, significant differences, second-order interactions, and higher-order effects, the assumption of additivity has been treated by variable rule users more as a working hypothesis than a formal null hypothesis. In the first versions of the variable rule program, no step-by-step procedure was available for carrying out a sequence of statistical evaluations and modifications of the null hypothesis. Instead, various heuristic and *ad hoc* approaches were used, as we will document below. The unfamiliarity of K & M with the role of statistical hypotheses in theory building is excusable, but their attempt to impute a rigid theoretical dogma to variable rule users is not, especially when this is contradicted by the spirit of that work and by explicit statements in it.

The notion of community grammar which K & M impute to users of variable rule methodology is not that suggested by any of the latter, who would agree with other sociolinguists on the definition of a speech community as a group of people who share a given set of norms of language: norms of referential interpretation as well as norms of social evaluation (Labov 1966: ch. 11; Hymes 1967). Variable rules are rules of production, and it is unfortunately true that very little of the work has been done which would establish the perceptual and evaluative correlates of the variation they record. We know that every speaker is a member of many nested and intersecting speech communities. We might be able to clarify the issue by asking whether the group of people whose speech production is described by a given set of variable rules share a uniform set of interpretative norms that respond to the variation in speech production so recorded. No subjective reaction study published so far has examined the perceptual correlates of internal constraints in a variable rule.[10]

Even on the level of speech production no user of variable rules has claimed, implicitly or explicitly, that a single rule per variable is always capable of accounting for the 'orderly heterogeneity' characteristic of a speech community (Weinreich, Labov & Herzog 1968). The very studies K & M cite as guilty of this assumption are rife with disclaimers, for example,

> '...this approach neatly solves the problem of community heterogeneity – perhaps too neatly; care should be taken to detect categorical rule differences where these exist... Further statistical methods must be developed in order to judge when small data sets on individual speakers can be aggregated without obscuring categorical distinctions between individual grammars [Cedergren & Sankoff (1974: 353)]

Furthermore, a great deal of the more recent research, available to K & M before the final versions of their text, has been preoccupied with the problem of community heterogeneity. Before going into this question in a substantive way, it will be necessary to examine more closely the critical assumption of independence of constraints. Then we can turn to the question of how uniform variable constraints have been found to be throughout the populations studied, and finally consider how variable rules deal with change and variation.

The independence of constraints

The assumption of the independence of variable constraints arises when we have at least two sets of cross-cutting environmental variants which condition the operation of the rule; as in the basic form of the *-t,d* deletion rule:[11]

[10] The original New York City subjective reaction test did compare the variable use of constricted [r] with categorical use (Labov 1972a: 149 ff.) but did not directly examine reactions to various degrees of variability. Tousignant (1978) elicited judgments of liaison use and omission in various syntactic contexts in Montréal French. This involved both normatively acceptable liaison and those particular to Québec varieties of French. Tousignant's data provide evidence for the interaction of 'grammaticality' judgments and salience. Generally, from context to context the rate of negative judgments to the omission of normatively prescribed liaison parallels the probability of application of the liaison rule. Conversely, the occurrences of improbable but normative liaison received higher proportions of positive judgments than more frequently occurring and hence less salient normative liaison. Lack of salience of non-normative but frequent liaison also tended to attenuate stigmatization reactions. A number of experiments have shown perceptual correlates of variation which allow us to infer indirectly some consequences of high vs. low levels of deletion for the retrievability of distinctions in perception and interpretation (Labov 1977; Torrey 1972; Biondi 1975).

[11] Angled brackets in the environment indicate features which favor the rule; the convention followed here is that factors with angled brackets are ordered in strength from top to bottom. For the sake of clarity, the least favored factor is also included in rule (3.1), although that is normally omitted as the residual case. A more systematic form of (3.1) would be

(3.1) $$t,d \rightarrow \langle \emptyset / \rangle C \left\langle \begin{matrix} \emptyset \\ \# \end{matrix} \right\rangle \_\_ \# \# \left\langle \begin{matrix} C \\ V \end{matrix} \right\rangle$$

That is, /*t*/ or /*d*/ is variably deleted at the ends of words after another consonant, more often if a consonant rather than a vowel follows, and more often if it is not separated from the preceding segment by a morpheme boundary than if it is (that is, if it is not a past-tense morpheme). The assumption of the independence of these two environmental constraints is equivalent to asserting that the realization of consonant or vowel in the following segment has exactly the same effect on the rule whether or not a grammatical boundary precedes the deletable element (that is, the /*t*/ or /*d*/ is a past-tense morpheme); and conversely, the effect of a past-tense boundary in constraining the rule is the same whether or not a consonant follows or a vowel. If this is so, then a single set of coefficients can be assigned to each of the variable constraints, and these coefficients, entered into the appropriate model, will predict closely the observed frequencies. It bears repeating that independence in this sense is the lack of interaction, or the additivity, of constraint effects in a model such as (2.3)–(2.5). It may well be interpretable in terms of probabilistic (or statistical) independence of random trials as dwelt upon by K & M and as we will discuss later, but this is not a necessary aspect of the analysis, and it requires further assumptions.

If the statistical fit of model and data is not good, we must be prepared to resolve the rule into its individual components:

(3.2) a. $t,d \rightarrow \langle \emptyset \rangle / C \_\_ \# \# C$

b. $t,d \rightarrow \langle \emptyset \rangle / C \_\_ \# \# V$

c. $t,d \rightarrow \langle \emptyset \rangle / C \# \_\_ \# \# C$

d. $t,d \rightarrow \langle \emptyset \rangle / C \# \_\_ \# \# V$

with unrelated probabilities for the four cases. If there is a good fit, we have justified the assembly of the individual rules into the rule schema indicated by the angled brackets.

Thus the great importance of the assumption of independence in variable rules is that it provides for the first time a way to test and justify the fundamental linguistic operation of writing abstract individual rule schemata and this test is based upon the characteristics of the data itself.

(3.1′) $$[-\text{cont}] \rightarrow \langle \emptyset \rangle / [+\text{cons}] \langle \emptyset \rangle \_\_ \# \# \langle +\text{cons} \rangle$$

This form opposes obstruents and liquids in the following environment to semi-vowels and vowels, which is the most significant division. A more detailed representation of the Philadelphia system shows

$$\left\langle \begin{matrix} +\text{seg} \\ +\text{cons} \\ -\text{voc} \end{matrix} \right\rangle$$

as the following environment, which indicates that the segments favoring deletion are ranked in the order obstruent, liquid, semi-vowel, vowel, and pause.

This does not mean that this assumption is expected to hold for any new case, or that variable rule analysis is based on the belief that a given set of internal constraints are independent. On the contrary, a variable rule program is a device for finding out if this is the case, and rather more can be learned about the grammar when it turns out not to be true. Cedergren's study of aspiration of Spanish /s/ (1973b) is a classic study of such lack of independence. The effect of determiner status on aspiration was not independent of whether or not the syllable was stressed; as a result, the rule for aspiration of stressed determiners had to be separated from the main rule for aspiration, a result which fitted well our understanding of the special semantic load carried by these pronouns.

Labov and Labov's study of the acquisition of the inversion rule by their daughter Jessie (1977) shows how the assumptions of the variable rule program are used to disprove the existence of rules as well as prove them. The analysis began with the assumption that early questions such as *What's this?* were produced by the adult rules of WH-fronting and inversion, as were such later questions of the form *Why did you do that?* It should be emphasized that the authors had no reason to believe that this was the case. They were interested in discovering the point in Jessie's development where the assumption did hold.

The successive variable rule analyses provided a number of grounds for rejecting the notion of an inversion rule in the early stages. One such indication was that in the early period, a preliminary variable rule analysis showed contraction favoring the putative inversion rule. Since contraction occurs at a much later stage in the derivation than an inversion rule, such an effect actually demonstrates that no such rule was applying, but that forms such as *What's this?* were

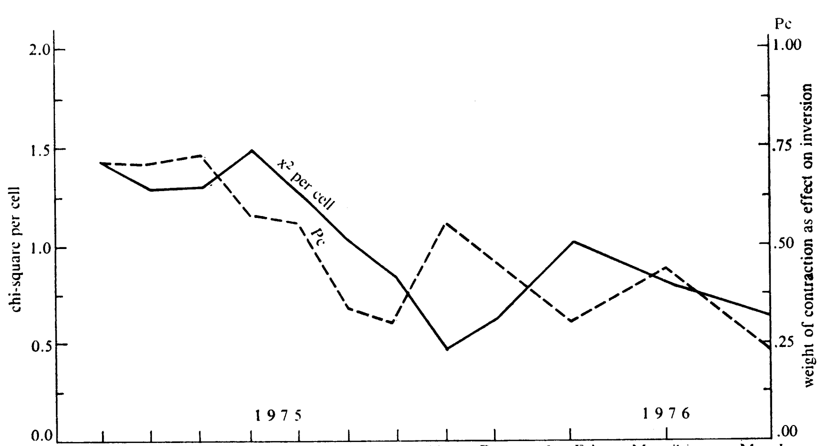

FIGURE 1. Improvement of fit of variable rule model to WH-inversion data compared to decrease in effect due to contraction.

produced directly by phrase-structure rules. At the point in time where this constraint was reversed, one could infer that the adult rules of WH-fronting and inversion were beginning to predominate in Jessie's grammar. Fig. 1 is based on a month-by-month variable rule VARBRUL 2 analysis of 8,500 questions asked by Jessie from 39 to 57 months of age. The steady decline in chi-square runs parallel to the decline in the effect of contraction. The fall in chi-square indicates an increasing fit of prediction and observation, and a confirmation that we are tracing the trajectory on which Jessie acquired the adult rule schema.

These results indicate that the power of variable rule analysis lies in the consequencies of disproving the assumption of independence of constraints as well as proving it. Such an assumption is not a belief about the data maintained in advance, but a device for discovering the nature of the phenomena in question. As such it corresponds to an assumption of non-interaction in ANOVA, and to a lack of interaction terms in multiple regression. Just as in these more classical statistical programs, once interaction is detected in a variable rule analysis, it can be evaluated, quantified and incorporated into the model. For example, if constraint a_k in the 'a' family is thought to interact with constraint b_j in the 'b' family, the model can account for this through the addition of a new constraint family, say the 'i' family, containing two terms: i_1, which represents the simultaneous presence of a_k and b_j, and i_2, which represents the absence of at least one of them. Equation (2.2) becomes

$$\text{(3.3)} \qquad \log \frac{p}{1-p} = \beta_0 + \beta_a + \beta_b + \beta_c + \ldots + \beta_n + \beta_i$$

In particular, in an environment containing a_k, b_j, c_h, ..., n_l,

$$\text{(3.4)} \qquad \log \frac{p}{1-p} = \beta_0 + \beta_{a_k} + \beta_{b_j} + \beta_{c_h} + \ldots + \beta_{n_l} + \beta_{i_1}$$

and in an environment which is identical except for a_r instead of a_k,

$$\text{(3.5)} \qquad \log \frac{p}{1-p} = \beta_0 + \beta_{a_r} + \beta_{b_j} + \beta_{c_h} + \ldots + \beta_{n_l} + \beta_{i_2}$$

Procedures and consequences of incorporating interaction and non-independence into variable rule methodology have been discussed by Sankoff (1977).

The relative uniformity of constraints

K & M's lack of familiarity with the main body of sociolinguistic research shows up most clearly in their reference to the 'large number of empirical studies that adopt [the variable-rule-as-community-grammar assumption] tacitly'. Far from adopting such an assumption, from the earliest pre-variable rule research to the present, quantitative studies of the speech community have carefully examined individual variation, and investigated the extent to which the same constraints influence the behavior of all individuals.

The original study of Martha's Vineyard (Labov 1963) was an examination of individual variation, and a search for the patterns that were common to all of the individuals recorded. The same holds for G. Sankoff & Cedergren's (1971) study of *l*-deletion in Montréal French, which was not phrased in terms of variable rules.

The New York City study began with an examination of style-shifting among individuals (Labov 1966, Ch. 4 = Labov 1972a, Ch. 3) and after presenting group means went on to give individual data for all five variables studied (Ch. 8, Figs 6, 7, 14, 15, 18, 19, 22, and Ch. 9, Figs 9, 10, 11). The variable rules presented for New York City in Weinreich, Labov & Herzog (1968: 170–6) are based on these individual data. Individual exceptions to the general patterns are considered in detail, for example, the case of Nathan B. (Labov 1966: 249–53). The extraordinary fact about the New York City situation was that almost all individuals responded to the same community norms in style-shifting and other measures of evaluation, even though they were stratified in production.[12]

G. Sankoff (1973, 1974) took pains to present individual data and discussed fully the methodological problems and epistemological questions in grouping speakers as if they shared constraints, versus trying to analyze them separately, with statistically inadequate data, as if they were assumed to share nothing.

The first reports of the South Harlem study, like many variable rule analyses, began with detailed analyses of a few individuals, reported first in Labov (1967) (= ch. 1, Labov 1972b). The variable rules reported in Labov *et al.* (1969) were based on the finding, rather surprising at the time, that all of the individuals studied showed the basic constraints of rule (3.1) above.

With the introduction of the variable rule program, it became possible to carry out more precise measurements of individual differences and to analyze the major constraints in finer detail. Thus the effect of a following consonant vs. vowel was resolved into the effect of following obstruent, liquid, glide, vowel, and pause (Labov 1975; Guy 1975). Guy's study of *-t,d* deletion in New York and Philadelphia, far from being the minor exception that it appears in a footnote added to the K & M discussion, is a report of a massive investigation of linguistic change and variation of the distribution of this rule across geographic, ethnic and age boundaries. Guy's examination of 14 constraints on *-t,d* deletion showed that

[12] The variable (oh) (the raising of the nucleus of *lost*, *office*, *law*, etc.) was an exception to this uniformity, since lower-class speakers did not show the characteristic style-shifting (see Fig. 5.2, Labov 1972a). As one of the more recent movements within the New York City system, (oh) had not yet been generalized to all social groups in 1963, and this would also be true for more recent changes such as the backing of (ay) and the fronting of (aw).

(1) The number of reversals of the expected ordering of constraint effects was inversely proportionate to the number of tokens measured, and for any comparison of environments with more than 30 tokens in the smaller cell, there were no reversals of the expected ordering.

(2) The patterns found in a very large amount of data from one Philadelphian (recorded throughout an entire day) were matched closely by 19 Philadelphians recorded for only one or two hours.

(3) The New York City and Philadelphia speech communities were sharply differentiated by the weight contributed by final pause to the probability of deletion: in the New York City, final pause promotes deletion like a consonant, while in Philadelphia it uniformly constrains deletion more than a vowel.[13]

A somewhat different tack was taken by Cedergren & Sankoff (1974) in their study of *r*-spirantization in Panamanian Spanish. Here a variable rule was calculated using independence of constraints and a division of the speakers into four socio-economic groups, so that each speaker was, in effect assigned one of four possible input parameters. (This exercise predated the high-capacity programs allowing a different parameter for each speaker.) Then each speaker's data with each constraint combination were compared with the predictions of the variable rule, using a chi-square test. Despite the fact that such a test tends to reject the null hypothesis far too often when there are less than five tokens of either variant (predicted) per comparison, a common situation with the *r* data and linguistic variation data in general, a large majority of the 79 speakers showed a statistically close fit of the rule and the data.

Rather than isolated exceptions, as K & M's footnoting would suggest, these examples show that the introduction of computer programs for calculating rules immediately led to an exploration of uniformity versus heterogeneity, and ways for testing for, estimating, and explaining heterogeneity when it exists. Indeed, the longstanding controversy over the relative homogeneity of individual rule forms, initiated by Bickerton (1971), has led to considerable methodological advance. In addition to the empirical studies of individuals and groups mentioned above, a more systematic approach to the problems of heterogeneity has culminated recently in a statistically powerful and original methodology, developed and programmed by Rousseau (1978) and also described by Rousseau & Sankoff (1978b). The idea here is to take data on a variable from a large number of individuals and to find the most likely way to divide these individuals into groups so that a single variable rule holds for the individuals within a group,

[13] Studies of the six communities in a 50-mile radius around Philadelphia indicate that the effect of final pause is geographically distributed along the North-Midland line. Communities closer to New York City show a relatively high weight contributed by final pause, close to the level of New York; those half-way between the cities show an intermediate value; and those fully in the Philadelphia area show a low value similar to Philadelphia (Labov 1977).

but these rules differ between groups. Neither the number of groups, the number of individuals per group, the social homogeneity or heterogeneity of speakers in a group, nor the nature of the differences or similarities in constraint effects are assumed *a priori*.

In a re-analysis of G. Sankoff, Kemp & Cedergren's (1978) data *ce que* vs. *qu'est-ce que* as the head of various complement clauses in Montréal French, Rousseau showed that a two-group analysis was significantly more likely than one done under the assumption of community homogeneity. In one group, which turned out to contain largely middle-class and bourgeois speakers using relatively little *qu'est-ce que* overall (input 0.13), the constraint which permitted the most *qu'est-ce que* was the indirect question environment, with a weight of 0.96, compared to 0.24, 0.36, and 0.19 for other types of clause. This is consistent with the extension of the direct question marker function of *qu'est-ce que* to indirect contexts. In the second group, which used much more *qu'est-ce que* overall (input 0.5), it is neither the indirect question context (weight 0.47) nor the superficially similar (post-verbal) headless relative construction (weight 0.46) which favors it, but rather a class of dislocated structures, with a weight of 0.81. This suggests certain normative influences have had an effect in favoring *ce que* among the latter group of speakers.

A similar reanalysis of data on the alternation of auxiliaries *avoir* and *être* in compound tenses of certain verbs, collected by G. Sankoff & Thibault (1977), showed that though some statistically significant grouping is detectable, the two groups found did not involve any dramatic constraint differences and the entire set of speakers could well be considered as homogeneous with respect to the relative tendencies of the different verbs to take *avoir*.

Rousseau applied the grouping procedure to Laberge's (1977) data on the variants *on* and *tu* (*vous*) as indefinite subject clitics, and discovered that the speakers fell into at least two and possibly three distinct groups. All three groups shared the same constraint pattern for a 'pragmatic' constraint family, in which a proverb-like, or moral, utterance strongly favors *on*, measured by a coefficient of 0.65 as compared to 0.35 for other uses of the indefinite. But a syntactic effect, involving the favoring of *tu* in pairs of implicationally related clauses, and the favoring of *on* in clauses imbedded in presentative heads is clearly neutralized among one group of speakers, and possibly even reversed among a small third group, though this latter grouping is of doubtful validity.

From various points of view, Rousseau's method seems the most appropriate way of dealing with heterogeneity within a community with respect to a variable. We have already mentioned an alternative method, the statistically more familiar method of simply adding interaction terms to the model, to take account of different weights given by different individuals to constraints in the same family. This latter method, however, is often impractical with linguistic variation data, especially when we wish to consider the data from each of a large number of individuals separately, rather than as lumped together according to some socio-demographic parameters. The large number of speaker parameters can give rise to a prohibitively large number of interaction parameters, particularly if a systematic search is to be made of possible interactions, and this will tend to overload the capacity of the computing system. Rousseau's method, on the other hand, while its search for possible groupings is systematic and exhaustive, does not require too many parameters – just one per linguistic constraint per group, and only one per individual speaker irrespective of the number of groups.

The two methods we have discussed for detecting and evaluating heterogeneity would seem to lead to different types of results. In the traditional statistical method, we obtain one equation, possibly including a number of interaction terms. With Rousseau's method, we obtain a number of equations, one per group, with no interaction terms necessary. Each of these forms has its advantages, but there is no logical difference, since either representation is mathematically convertible into the other.

Change and variation in the speech community

We have seen that the notion that the uniformity of variable rule patterns is 'tacitly assumed' in sociolinguistic studies is based on a lack of familiarity of K & M with those studies. The assumption 'that there is no interaction between linguistic and social constraints' is almost entirely due to K & M, and has little relation to the theory and practice of variable rule analyses.

So far, we have been discussing the independence and relative uniformity of internal linguistic constraints. At the outset, it was noticed that this relative uniformity applies only to the direction of the major constraints, and less attention was paid to the relative ordering of those constraints, and even less to their precise strength. Studies of the contraction and deletion of the copula within the Jets and Cobras then showed that the central core could be differentiated from the secondary and peripheral members and from those outside the groups ('lames') (Labov 1973). Other studies of *-t,d* deletion, cited by K & M, show shifts in the ordering of constraints that differentiated speakers according to age and social allegiance.

Here we are dealing with interaction between a social fact – age – and the internal linguistic constraints. Though their ordering can be established for the entire community as a whole, their values and relative strengths cannot be stated independently of the age of the speaker.

The most dramatic example of the interaction of age and internal constraints arises in the most recent analysis of the 'ambiguous' clusters in *lef+t*, *kep+t*, *tol+d*, etc. Here the + boundary registers the existence of a derivational suffix, which is opposed to the inflectional suffix in *roll#d* by (1) regressive assimilation, (2) lack of productivity, and (3) change in stem form. The process of *-t,d* deletion allows us to see whether speakers treat this final /t/ like the past-tense morpheme or like monomorphemic /t/ in *fist*, and so registers the degree to which they have analyzed the data into a stem alternation /liyv ~ lef/, etc., with a /+t/ suffix, or as a suppletive alternation /liyv ~ left/, etc.

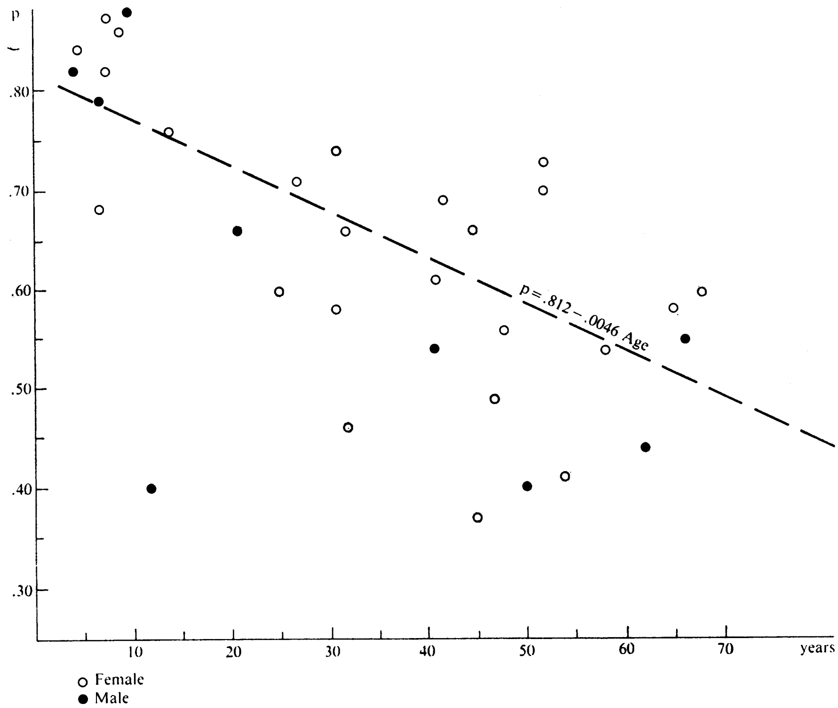

FIGURE 2. Relation between age and probability of deleting /t/ in ambiguous clusters across a derivational boundary in *left*, *kept*, etc.

Fig. 2 shows the result of 33 variable rule analyses of individual Philadelphians, carried out by Sally Boyd of LCV and incorporating Guy's analyses. The vertical axis represents the weight contributed by the presence of an ambiguous cluster to the probability of *-t,d* deletion. The horizontal axis shows the age of the speakers, from 5 to 69 years of age. At the outset, the ambiguous clusters favor deletion as much or more than monomorphemic forms; with advancing age, ambiguous clusters begin to constrain the rule, approaching the value of the past-tense morpheme. The coefficient of the regression line fitted to the data indicates that with each year of age, the value of the ambiguous cluster constraint is expected to fall by 0.0046.

This result demonstrates that speakers analyze the derivational morphology of past-tense clusters more deeply as they grow older. There is no reason to believe that the increasing depth of analysis of derivational morphology is confined to this one case. Such a change in age has important consequences for our views of all derivational morphology and phonological rules that depend upon morphological boundaries.

The *ce que/qu'est-ce que* variable discussed in section 3.2 also shows a complicated interaction between age, social class and linguistic constraint. Both G. Sankoff, Kemp & Cedergren (1978) and Rousseau & Sankoff (1978b) find a greater differential between indirect questions and headless relatives, in their effect on *qu'est-ce que* usage, among middle-class or bourgeois vs. working-class speakers, and this appears tied to the different mechanisms with which younger speakers in the opposing classes are converging to a common rule for this variable.

There is clearly a great difference between linguistic and extralinguistic environments as far as the degree of independence of factors is concerned. Internal linguistic factors are typically independent of each other, and this fact provided much of the motivation for the development of variable rules. The advantages of this working assumption for the investigation of linguistic structure have been outlined in the previous section. External, social constraints more frequently show interaction in their relation to each other and to linguistic variables, as with the examples that K & M discuss. The variable rule program treats them both in the same statistical way, although in the first case we can expect the assumption to be satisfied more often than not, in the second case the reverse.

There is no reason to be alarmed at such a situation. There may be some relation between the value of a working hypothesis and how often that hypothesis is found to be correct, though statistical null hypotheses are more frequently straw men set up to justify more elaborate hypotheses. The variable rule program is designed with the characteristics of internal linguistic constraints in mind: their skewed distributions as well as their characteristic independence. If we were to deal only with external constraints, the variable rule program might not be our choice for multivariate analysis. There would be no reason not to fill all cells more or less evenly (as with the sample of Montréal French speakers, D. Sankoff & G. Sankoff 1973, or Summerlin 1972), transform the data of form A/T (see p. 195 above) according to some function such as in (2.3)–(2.5), and then apply a standard multivariate analysis program like ANOVA which requires evenly distributed data, but which handles interaction in a very detailed and rigorous way.

At present, a linguistic analyst has a wide variety of options in approaching external linguistic variables. By incorporating them into a variable rule program, one runs little risk of losing the precision of the analysis of internal variables. As in (3.3), simple cases of interaction can be measured by adding an additional factor group which represents the interaction of the two categories concerned. For example, to account for the typical acceleration of style-shifting on the part of female, lower-middle-class speakers, one can add a factor group which registers the co-presence or absence of female gender and lower-middle-class status, in addition to the sex and class groups already present. On the other hand, one can run separate variable rule programs for subgroups or for individuals as in Fig. 2, or for stages in the acquisition of language by one individual, as in Fig. 1, or even find the relevant subgroups in the population using Rousseau's method.

When it comes to the writing of the variable rules, the presence of interaction does create a problem. As pointed out above, the absence of independence of the internal constraints is justification for writing separate rules. It follows that interaction of social constraints with internal constraints and with each other makes it less meaningful to simply add on sex, class, or ethnicity as 'wider' constraints on a variable rule such as (3.1).

Having documented the capacity of variable rules to study interaction, we should point out some recent results indicating a tendency towards independence of linguistic and social constraints for a certain range of variables.

In Weiner and Labov's investigation of the choice between agentless passive and active with generalized pronouns (p. 199), a number of strong internal constraints were derived through cross-tabulations and VARBRUL 2. A stylistic factor was located, which might indeed have been stronger if more formal styles were investigated. But social factors such as sex, class and ethnicity, which might have been expected to influence the choice, proved to have very little effect. More importantly, the entire set of external factors, including style and age, remained almost invariant under the most radical re-organizations of the internal factors. Table 2 demonstrates this independence of internal and external constraints on the passive rule by showing the entire set of changes in factor values as individual factors are successively eliminated in the log-likelihood significance test (p. 198–200).

Thus independence from social constraints takes two forms: first, a reduction in the types and strengths of social factors which influences the output of the rule itself, and secondly, an absence of interaction with internal constraints.

It is likely that we will encounter this situation with increasing frequency as we examine higher alternations in the grammar. Emirkanian (1978) has studied gapping and other conjunction-reduction phenomena in French. These high-level syntactic processes vary widely amongst themselves, but showed relatively less socially conditioned variability among the diverse speakers Emirkanian sampled.

The strongest social effect appears to be attributed to the words and sounds of the language – that is, the output of the rule system. The more abstract the variation, the less apt we are to find social influences exerted upon it.

4. THE RELATION OF VARIABLE RULES TO GENERATIVE GRAMMAR

At both the beginning and the end of K & M's review, they present their view

TABLE 2. *Significance of constraints on the passive using log-likelihood differences*

Run	p0	Given					Parallel S.S.					Pre-p.		Style		Sex		Class			Age		Log Like. Chi-square
		0	1	2	3	4	0	1	2	3	4	Yes	No	C	I	M	F	W	M	B	K	A	
Elimination of external variables																							
1	0.74	0.40	0.39	0.58	0.61	0.53	0.21	0.45	0.63	0.67	0.56	0.69	0.31	0.54	0.46	0.50	0.50	0.58	0.46	0.45	0.46	0.54	—
2	0.74	0.40	0.39	0.58	0.61	0.53	0.21	0.45	0.63	0.67	0.56	0.69	0.31			0.49	0.51	0.58	0.46	0.45	0.46	0.54	−5.57*
3	0.74	0.40	0.39	0.58	0.61	0.53	0.21	0.45	0.63	0.67	0.57	0.69	0.31	0.54	0.56			0.58	0.46	0.45	0.46	0.54	0.00
4	0.75	0.40	0.39	0.58	0.61	0.53	0.21	0.44	0.63	0.67	0.57	0.69	0.31	0.54	0.46	0.50	0.50	0.54		0.46	0.44	0.56	−9.65**
5	0.74	0.40	0.39	0.58	0.61	0.53	0.22	0.45	0.63	0.67	0.56	0.69	0.31	0.54	0.46	0.49	0.51	0.55	0.45		0.43	0.57	−4.84*
6	0.76	0.41	0.39	0.58	0.61	0.53	0.21	0.45	0.63	0.67	0.56	0.69	0.31	0.54	0.46	0.50	0.50				0.43	0.57	−11.54**
7	0.75	0.41	0.38	0.58	0.62	0.53	0.31	0.45	0.63	0.67	0.56	0.68	0.32	0.53	0.47	0.59	0.51	0.61	0.47	0.42			−2.05
Elimination of internal variables																							
8	0.61	0.41	0.39	0.58	0.61	0.52	0.21	0.45	0.64	0.64	0.58			0.54	0.46	0.50	0.50	0.57	0.47	0.46	0.46	0.54	−55.04***
9	0.72	0.40	0.40	0.58	0.61	0.52	0.24	0.48	0.60	0.65		0.69	0.31	0.54	0.46	0.50	0.50	0.58	0.46	0.46	0.46	0.54	−0.85
10	0.67	0.40	0.40	0.58	0.61	0.52	0.28	0.57	0.70			0.69	0.31	0.54	0.46	0.50	0.50	0.58	0.46	0.45	0.46	0.54	−0.87
11	0.63	0.40	0.40	0.58	0.62	0.52	0.34	0.66				0.69	0.31	0.54	0.46	0.50	0.50	0.58	0.46	0.46	0.46	0.54	−10.41**
12	0.51	0.36	0.65	0.65	0.60	0.50						0.69	0.31	0.54	0.46	0.49	0.51	0.57	0.47	0.48	0.46	0.54	−71.61***
13	0.73	0.42	0.41	0.60	0.58		0.21	0.45	0.63	0.67	0.56	0.69	0.31	0.54	0.46	0.50	0.50	0.58	0.46	0.46	0.46	0.54	−0.88
14	0.70	0.45	0.44	0.61			0.22	0.44	0.63	0.67	0.56	0.69	0.31	0.54	0.46	0.50	0.50	0.58	0.46	0.45	0.46	0.54	−1.13
15	0.63	0.46	0.54				0.27	0.43	0.61	0.65	0.54	0.68	0.32	0.54	0.46	0.50	0.50	0.58	0.46	0.45	0.46	0.54	−18.41***
16	0.66						0.25	0.44	0.62	0.66	0.55	0.69	0.31	0.55	0.45	0.50	0.50	0.58	0.46	0.51	0.46	0.54	−25.27***

Given: 0 = not given; 1–4 = coreferent in *i*th preceding clause
Parallel S.S.: 0 = no parallel; 1–4 = length of preceding string of clauses w/ parallel subjects
Pre-p.: preceding passive
C = careful; I = casual
M = male; F = female
W = white working; M = white middle; B = black working
K = adolescents; A = adults
*$p < 0.05$
**$p < 0.01$
***$p < 0.001$

that variable rules do in fact represent a radical departure from generative theory; that the use of the generative format is a 'premature rush to formalism', and that we must await a 'more fully formal theory of variability in token formation', presumably quite distinct from a theory about type formation.

K & M have a right to their opinion on all these points. Moreover, it is plainly not their intent to disparage current work on variation, but rather to encourage it. At the same time, their assessment of the relation to generative grammar is based on a rather narrow view as to what that theory has contributed to the understanding of language. It also ignores serious efforts on the theoretical level to relate the study of variation to the generative approach. K & M appear to set considerable store upon the idea that generative grammar describes competence and not performance; they also put great weight on the argument that generative theory deals only with types, and cannot make contact with data based on what people actually say. Despite their disclaimer about not relying on 'linguistic autonomy' and their footnote 13 acknowledging trends towards studying usage within generative grammar, K & M continuously fall back on the more ideological aspects of early generativism which do not really have any serious connection with the important contributions that Chomsky and his school have made to our understanding of language. Among others, these have to do with the systematic investigation of relations between sentences, the willingness to consider more abstract hierarchical structures in syntax and phonology, and the stress upon the importance of language acquisition for the development of the theory.

Let us, however, consider for a moment the purely formal aspects of generativism. True, grammars are finite devices for generating infinite sets called languages. But probabilistic grammars, while of course having a different logical status, do not constitute a 'contrary' nor even a 'radical' departure as K & M claim. Mathematically, and in terms of model construction, probabilistic grammar is a natural and easy extension of ordinary grammar rather than a 'patch' or 'graft' as K & M phrase it. Indeed this extension has been discovered and rediscovered many times by both mathematicians and linguists, going back at least to the mid-1960s. For the type of context-free grammar K & M use to illustrate the generative principle, there is a substantial and interesting literature on both the mathematical and linguistic aspects of probabilistic extensions, for example, S. Klein (1965), Grenander (1967), Kherts (1968), Horning (1969), Suppes (1970), D. Sankoff (1971, 1972), Soule (1974), W. Klein (1974), Heidelberger Forschungsprojekt 'Pidgin-Deutsch' (1975, 1976, 1977, 1978), and D. Sankoff (1978b). That these developments may not have been foreseen when the early notions of generativity were evolving does not bear on their mathematical correctness or their appropriateness as models for linguistic competence or performance.

Again on the purely formal level, the same can be said for variable rules. Here it is transformational grammar which is probabilized and the focus is on a single rule acting on a syntactic or phonological entity. The probabilistic extension is two-fold. As with probabilistic context-free grammars, the notion of rule probability is postulated. In addition, the generative notions of rule constraints are considered to be extreme cases, categorical or qualitative instances of a more general quantitative functioning of constraints. It is this, together with the idea that cross-cutting constraint effects combine in a systematic way, which is the mathematical basis of variable rule theory. This is a natural and economical generalization of the generative concept of rule, even if as a 'logical' object it necessarily has different status. Grammars containing probabilistic rules define membership in a language set just as ordinary grammars do, but they also predict frequencies of sentences, and more important, of various sentence types. The naturalness of the extension is further attested by the ease with which the generative rule notation was adapted to account for variability.

On the formal level, then, and this is argued in more detail by D. Sankoff (1978b), probabilistic extensions of generative theory are neither radical, premature, mistaken, nor contradictory. Turning to the appropriateness of this extended formalism as a theory of linguistic performance, we have already shown how K & M's critiques of statistical models and procedures are uninformed and incorrect. It is true, however, that the model represented by (2.1) and in turn those implicit in (2.3) and (2.4) were not at first seen in the context of their being special cases of the class of transformed-additive models, and that hence postulates were made about the mathematical form of independence of constraints that turned out to be overly specific. The current use of (2.5) stems from statistical and data-analytic considerations as being a more generally applicable model, rather than any more detailed an interpretation in terms of probabilistic mechanisms. Indeed, the discussion of probabilistic independence connected with (2.3) and (2.4) contained in Cedergren & Sankoff (1974) was phrased to serve an expository function for readers unfamiliar with probability theory. Probabilistic notions of causality were introduced, but nowhere were the probability calculations identified with psycholinguistic or neurolinguistic mechanisms nor was the specific existence of such mechanisms postulated. The emphasis on psychological interpretations and difficulties therewith is entirely due to K & M themselves, and contradicts their own admission that:

> Strong claims about the nature of the mathematical complexity of the mental-neural abilities that underly 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, panlectal grammars, or what have you.

Thus as an extension to generative theory, variable rules hardly warrant the criticism advanced by K & M. Their own tendencies toward psychological and neurological reductionism combined with their arbitrary treatment of distinctions between theory-building, modeling, and statistical inference, as well as their mathematical errors, seem to have led them to an unjustified attack on the theoretical status of variable rules.[14]

More substantively, as we noted in section 1, an early motivation for a connection with generative grammar lay in the relationship of the contraction and deletion rules with the stress assignment and vowel reduction rules of Chomsky & Halle (1968). It has also been evident from the start that variable rules were the appropriate mechanism to describe the acquisition of grammar. The recent studies of the acquisition of the inversion rule (see p. 205) give us a direct view of the rule-forming capacity of a child (Labov & Labov 1977). In the analysis of variation in the inversion rule over two and a half years and 20,000 questions, we obtain a vivid view of the formation of rule schemata, in particular the gradual integration of questions formed with the eight WH-forms into a single set of rules for WH-attachment, WH-fronting, and auxiliary inversion.

Though formalisms are necessary to focus an analysis on particular linguistic relations, it would be foolhardy to put too much stock on any one formalism. The theory that we are constructing is not a new form of model-building, and we do not make the error of confusing the set of rules we write with the grammatical processes that people use.

Indeed, the notational and formal questions surrounding variable rules have receded in importance, much as have analogous questions in many areas of linguistic theory. Though the methodology of variable rules was motivated by and developed in conjunction with the project to incorporate variability in generative grammar, it would be a mistake to think that this methodology is logically tied to a particular grammatical formalism, or a particular domain of grammar such as phonology or morphology. Whenever a choice process is postulated in linguistic performance, especially choice which is conditioned by a number of cross-cutting linguistic and/or extralinguistic factors, a variable rule analysis, which is after all a statistically general way for handling conditioned binomial variables of all types, can be fruitfully undertaken. This applies, for

[14] In their footnote 1, K & M mention the implicational scale as the major analytic device competing with variable rules in the analysis of linguistic variability. Variable rule users have long been convinced that these two modes of analysis are, for purposes of organizing variation data, largely equivalent (Fasold 1975; D. Sankoff & Rousseau 1974). Of course, a variable rule analysis does make stronger claims about the relationships between the probabilities generating the data in different contexts, while a scaling analysis requires stronger relations of ordering among the variant frequencies themselves. In particular, and Kay (1978) goes on at length about this, scales generally involve many contexts in which one variant appears 100% of the time and the other variant never, and it has been thought impossible for this type of categorical behavior to be represented in terms of variable rules. Recently, however, Rousseau (1978, ch. 4) has discovered that the mathematics of so-called knockout factors (Labov 1969) and technical knockout factors (Rousseau & Sankoff (1978b)) coincides completely with the implicational scale structuring of categorical versus variable contexts (see Rousseau & Sankoff 1978c). This completely invalidates Kay's argument.

example, to the choice between lexical items which might be applied to the same referent (Labov 1978). What is of special interest is that it applies to the alternation between surface variants to give meaningful and useful results which can then apply to the further analysis of the more abstract grammatical processes. A prime example is the *ce que/qu'est-ce que* data, where the interaction of social and linguistic constraints was firmly established and understood without prejudging the problem of the underlying grammatical relationship of the variants.

One consequence of the historical association of variable rule analysis with generative grammar is that it has focused on counts of rule applications versus non-applications, which only produce binary or binomial data. Other phenomena, such as synonym selection, and other ways of looking at variability, might involve, at a single step, the choice between three or more variants. This leads to multinomial data. Instead of A applications and $T-A$ non-applications out of T instances of the variable in a given context, we might have A tokens of the first variant, B tokens of the second, C of the third, and so on, where

(4.1) $$A+B+C+\ldots = T$$

There are many existing data sets which might be (and are being) re-analyzed from this point of view; Laberge's *on/tu-vous* data is an obvious example. Instead of considering *tu* and *vous* as essentially the same variant opposed to *on*, they could be considered separate variants in a trinomial model. Cedergren had to postulate a specific series of reduction rules for each of /s/ and for /r/ in Panamanian Spanish before she could apply a variable rule analysis. It might well be revealing to analyze these data according to a multinomial model before making any linguistic assumptions.

What is the multinomial generalization of the variable rule model? Instead of application and non-application probabilities p and $1-p$, respectively, we postulate $p, q, r, \ldots, z$ such that

(4.2) $$p+q+r+\ldots+z = 1$$

A model similar to (2.5) is still applicable but we now require several such equations instead of just one as in the binomial case; the trinomial case becomes

(4.3) $$\log \frac{p}{q} = \beta_0+\beta_a+\ldots+\beta_n$$

(4.4) $$\log \frac{q}{r} = \gamma_0+\gamma_a+\ldots+\gamma_n$$

and a third equation is implied by the above two:

(4.5) $$\log \frac{r}{p} = \delta_0+\delta_a+\ldots+\delta_n$$

where

(4.6) $$\delta_i = -\beta_i - \gamma_i$$

Again, constraints analogous to (2.6) should be added to the model to ensure uniqueness of the parameter estimates; these have the form

(4.7) $$\sum_i \beta_{ai} = \sum_j \beta_{bj} = \ldots = \sum_k \beta_{nk}$$
$$= \sum_i \gamma_{ai} = \sum_j \gamma_{bj} = \ldots = \sum_k \gamma_{nk}$$
$$= \sum_i \delta_{ai} = \sum_j \delta_{bj} = \ldots = \sum_k \delta_{nk}$$

Computational procedures for multinomial models exist (Jones 1975), but they are not efficient enough to handle the type of data sets which arise in linguistic variation studies. New methods are currently under development (D. Sankoff 1978c).

To conclude, the concept of variable rule has notational (formal), theoretical, data-analytic and substantive empirical aspects. Its evolution through various empirical studies as well as its notation has been largely tied to specific questions within the generative approach to grammar. The results of these studies have been extensions, improvements and clarifications of knowledge gained through, or accessible through, previous qualitative work. On the theoretical level, variable rules as probabilistic models reflect and render mathematically and logically rigorous this extension or generalization of grammar to include probability as well as possibility. This aspect is not, of course, specific to generative grammars although we argue (against K & M) that it is eminently meaningful for that model. It is pertinent in any grammatical theory which incorporates choice as a mechanism for theoretically relating and differentiating sentences. So, of course, are the data-analytic aspects, the methods for gathering data, statistically fitting and testing the models. This attitude to linguistic analysis is not limited to the range of variables discussed here, but is being extended on the one hand to the mechanism of sound change and on the other to increasingly more abstract areas of phonological and grammatical variation. In this fashion, we hope to move steadily from the known to the unknown, deriving principles of increasing generality, using the insights of generative grammar wherever helpful without being governed by its dogma.

REFERENCES

Bailey, B. (1969). Toward a new perspective in Negro English dialectology. *American Speech*, **40.** 171–7.

Baugh, J. (to appear). A re-examination of the black English copula. In W. Labov (ed.), *Quantitative analyses of linguistic structure.* New York: Academic Press.

Bickerton, D. (1971). Inherent variability and variable rules. *Foundations of Language* **7.** 457–92.

—— (1973). Quantitative versus dynamic paradigms: The case of Montreal *que*. In C.-J. N. Bailey & R. W. Shuy (eds), *New ways of analyzing variation in English*. Washington, D.C.: Georgetown University Press. 23–43.
Biondi, L. (1975). *The Italian-American child: His sociolinguistic acculturation*. Washington D.C.: Georgetown University Press.
Cedergren, H. J. (1973a). Interplay of social and linguistic factors in Panama. Ph.D. thesis, Cornell University.
—— (1973b). On the nature of variable constraints. In C.-J. N. Bailey & R. W. Shuy (eds), *New ways of analyzing variation in English*. Washington, D.C.: Georgetown University Press. 13–27.
—— & D. Sankoff (1974). Variable rules: Performance as a statistical reflection of competence. *Language*, **50**. 333–55.
Chomsky, N. & M. Halle, (1968). *The sound pattern of English*. New York: Harper & Row.
Cox, D. R. (1970). *The analysis of binary data*. London: Methuen.
Emirkanian, L. (1978). Phénomènes de coordination et de réduction dans la langue parlée. Ph.D. thesis, Université de Provence.
Fasold, R. W. (1972). *Tense marking in black English: A linguistic and social analysis*. Arlington: Center for Applied Linguistics.
—— (1975). The Bailey wave model: A dynamic quantitative paradigm. In R. W. Fasold & R. W. Shuy (eds), *Analyzing variation in language*. Washington, D.C.: Georgetown University Press. 27–58.
—— (1978). Language variation and linguistic competence. In D. Sankoff (ed.), *Linguistic variation: Models and methods*. New York: Academic Press. 85–95.
Gazdar, G. (1976). Quantifying context. *York Papers in Linguistics*, **6**. 117–29.
Grenander, U. (1967). Syntax-controlled probabilities. Technical report, Brown University, Division of Applied Mathematics.
Guy, G. R. (1975). Variation in the group and the individual: The case of final stop deletion. *Pennsylvania Working Papers on Linguistic Change and Variation* **1**(4). Philadelphia: U.S. Regional Survey.
Haberman, S. J. (1974). *The analysis of frequency data*. University of Chicago Press.
Heidelberger Forschungsprojekt 'Pidgin-Deutsch' (1975). *Sprache und Kommunikation ausländischer Arbeiter, Analysen, Berichte, Materialien*. Kronberg: Scriptor Verlag.
—— (1976). Zur Sprache ausländischer Arbeiter: Syntaktische Analysen und Aspekte des kommunikativen Verhaltens. *Zeitschrift für Literaturwissenschaft und Linguistik*, **18**. 78–121.
—— (1977). Transitional grammars in the acquisition of German by Spanish and Italian workers. In J. M. Meisel (ed.), *Langues en contact–Pidgins–Creoles–Languages in contact*. Tübingen.
—— (1978). The acquisition of German syntax by foreign migrant workers. In D. Sankoff (ed.), *Linguistic variation: Models and methods*. New York: Academic Press. 1–22.
Horning, J. J. (1969). A study of grammatical inference. Technical Report CS 139. Stanford University, Computer Science Department.
Hymes, D. (1967). Models of the interaction of language and social setting. *Journal of Social Issues*, **23** (2). 8–28.
Jones, R. H. (1975). Probability estimation using a multinomial logistic function. *Journal of Statistical Computation and Simulation*, **3**. 315–29.
Kay, P. (1978). Variable rules, community grammar, and linguistic change. In D. Sankoff (ed.), *Linguistic variation: Models and methods*. New York: Academic Press. 71–83.
—— & McDaniel (1979) On the logic of variable rules. *LinS*, **8**(2), 151–87.
Kherts, M. M. (1968). Entropy of languages generated by automated or context-free grammars with a single-valued derivation. *Nauchno-Tekhnicheskaia Informatsia, Series* **2**. 29–34.
Klein, S. (1965). Control of style with a generative grammar. *Language*, **41**. 619–31.
Klein, W. (1974). *Variation in der Sprache. Ein Verfahren zu ihrer Beschreibung*. Kronberg/Ts: Scriptor Verlag.

Kruskal, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society* (*B*), **27**. 251–63.
Laberge, S. (1977). Etude de la variation des pronoms sujets définis et indéfinis dans le français parlé à Montréal. Ph.D. thesis, Université de Montréal.
Labov, W. (1963). The social motivation of a sound change. *Word*, **19**. 273–309.
—— (1966). *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.
—— (1967). Some sources of reading problems for Negro speakers of non-standard English. In A. Frazier (ed.), *New directions in elementary English*. Champaign, Ill: National Council of Teachers of English. 140–67.
—— (1969). Contraction, deletion, and inherent variability of the English copula. *Language*, **45**. 715–62.
—— (1972a). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
—— (1972b). *Language in the inner city*. Philadelphia: University of Pennsylvania Press.
—— (1973). The linguistic consequences of being a lame. *LinS*, **2**. 81–115.
—— (1975). The quantitative study of linguistic structure. In K.-H. Dahlstedt (ed.), *The Nordic languages and modern linguistics*. Stockholm: Almqvist & Wiksell. 188–244.
—— (1977). Categorical discrimination along a new phoneme boundary. 52nd LSA meeting, Chicago.
—— (1978). Denotational structure. Papers from the 14th regional meeting of the Chicago Linguistic Society.
——, Bower, A., Hindle, D., Dayton, E., Lennig, M. & Schiffrin, D. (1979). *Linguistic change in Philadelphia*. Technical Progress Report on NSF Grant SOC 7500245 Philadelphia: U.S. Regional Survey.
——, Cohen, P., Robins, C. & Lewis, J. (1968). *A study of the non-standard English of Negro and Puerto Rican speakers in New York City*. Philadelphia: U.S. Regional Survey.
—— & Labov, T. (1977). Learning the syntax of questions. In R. Campbell & P. Smith (eds), *Recent advances in the psychology of language*. New York, Plenum Press. Also as: Das Erlernen der Syntax und von Fragen. *Zeitschrift für Literaturwissenschaft und Linguistik 23/24*. Also as: L'apprentissage de la syntaxe des interrogations. *Langue Française*, **34**. 52–80.
Legum, S., Pfaff, C., Tinnie, G. & Nicholas, M. (1971). The speech of young black children in Los Angeles. Technical report 33, Southwest Regional Laboratory.
Lemle, M. & Naro, A. J. (1977). Competências basicas do Portugues. Report, Rio de Janeiro.
Lindsey, J. K. (1975). Likelihood analysis and tests for binary data. *Applied Statistics*, **24**. 1–16.
Lüdicke, A. (1977). Untersuchungen zum Gebrauch der Verneinungspartikel *ne* im gesprochenen Französich. Master's thesis, University of Hamburg.
Mitchell-Kernan, C. (1969). Language behavior in a black urban community. Working paper 23. Berkeley: Language Behavior Research Laboratory.
Naro, A. J. & Lemle, M. (1976). Syntactic diffusion. *Papers from the parasession on diachronic syntax*. Chicago Linguistic Society. 221–47.
Rickford, J. (1975). Carrying the new wave into syntax: The case of black English BIN. In R. W. Fasold & R. W. Shuy (eds), *Analyzing variation in language*. Washington, D.C.: Georgetown University Press. 162–83.
Rousseau, P. (1978). Analyse de données binaires. Ph.D. thesis, Université de Montréal.
—— & Sankoff, D. (1978a). Advances in variable rule methodology. In D. Sankoff (ed.), *Linguistic variation: Models and methods*. New York: Academic Press. 57–69.
—— & —— (1978b). A solution to the problem of grouping speakers. In D. Sankoff (ed.), *Linguistic variation: Models and methods*. New York: Academic Press. 97–117.
—— & —— (1978c). Singularities in the analysis of binomial data. *Biometrika*, **65**.
Sankoff, D. (1971). Branching processes with terminal types: Application to context-free grammars. *Journal of Applied Probability*, **8**. 233–40.
—— (1972). Context-free grammars and nonnegative matrices. *Linear Algebra and its*

Applications, **5**. 277–81.
—— (1975). VARBRUL version 2. Unpublished program and documentation.
—— (1977). Statistical dependence and interaction within the variable rule framework. 39th LSA summer meeting, Honolulu.
—— (1978a). *Linguistic variation: Models and methods*. New York: Academic Press.
—— (1978b). Probability and linguistic variation. *Synthèse*, **37**. 217–38.
—— (1978c). Réduction d'un modèle logistique-linéaire de données multinômiales. 46th Congress of L'association canadienne-française pour l'avancement des sciences.
—— & Laberge (1978). The linguistic market and the statistical explanation of variability. In D. Sankoff (ed.), *Linguistic variation: Models and methods*. New York: Academic Press. 239–50.
—— & Rousseau, P. (1974). A method for assessing variable rule and implicational scale analyses of linguistic variation. In J. L. Mitchell (ed.), *Computers in the Humanities*. Edinburgh University Press. 3–15.
—— & Sankoff, G. (1973). Sample survey methods and computer-assisted analysis in the study of grammatical variation. In R. Darnell (ed.), *Canadian languages in their social context*. Edmonton: Linguistic Research Inc. 7–64.
Sankoff, G. (1973). Above and beyond phonology in variable rules. In C.-J. N. Bailey & R. W. Shuy (eds), *New ways of analyzing variation in English*. Washington, D.C.: Georgetown University Press. 44–61.
—— (1974). A quantitative paradigm for the study of communicative competence. In R. Bauman & J. Sherzer (eds), *Explorations in the ethnography of speaking*. Cambridge University Press. 18–49.
—— & Cedergren, H. J. (1971). Some results of a sociolinguistic study of Montreal French. In R. Darnell (ed.), *Linguistic diversity in Canadian society*. Edmonton: Linguistic Research Inc. 61–87.
——, Kemp, W. & Cedergren, H. J. (1978). The syntax of *ce que/qu'est-ce que* variation and its social correlates. In R. W. Shuy & J. Firsching (eds), *Dimensions of variability and competence*. Washington, D.C.: Georgetown University Press.
—— & Thibault, P. (1977). L'alternance entre les auxiliaires *avoir* et *être* en français parlé à Montréal. *Langue Française*, **34**. 81–108.
Soule, S. (1974). Entropies of probabilistic grammars. *Information & Control*, **25**. 57–74.
Stewart, W. (1970). Toward a history of Negro dialect. In F. Williams (ed.), *Language and poverty*. Chicago: Markham. 351–79.
Summerlin, N.-J. (1972). A dialect study: Affective parameters in the deletion and substitution of consonants in the Deep South. Ph.D. thesis, Florida State University.
Suppes, P. (1970) Probabilistic grammars for natural languages. *Synthèse*, **22**. 95–116. Also in D. Davidson & G. Harmon (eds), (1972). *Semantics of natural language*. Dordrecht: Reidel 741–62.
Torrey, J. W. (1972) The language of black children in the early grades. Connecticut College, Department of Psychology Report.
Tousignant, C. (1978). La liason consonantique en français montréalais. Master's thesis, Université de Montréal.
Weiner, E. J. & Labov, W. (1977). Constraints on the agentless passive. 39th LSA summer meeting, Honolulu.
Weinreich, U., Labov, W. & Herzog, M. (1968). Empirical foundations for a theory of language change. In W. Lehmann & Y. Malkiel (eds), *Directions for historical linguistics*. Austin: University of Texas Press. 97–195.
Wolfram, W. (1969). *A sociolinguistic description of Detroit Negro speech*. Washington, D.C.: Center for Applied Linguistics.
—— (1974). *Sociolinguistic aspects of assimilation: Puerto Rican English in New York City*. Arlington: Center for Applied Linguistics.
Zwicky, A. (1970). Auxiliary reduction in English. *Linguistic Inquiry*, **1**. 323–36.