A TEST FOR MIXED RULES

DAVID SANKOFF AND PASCALE ROUSSEAU

A recurrent dilemma in the quantitative assessment of the contextual factors influencing the variable presence of a linguistic form is the question of whether *absence* results from a *deletion* process or, rather, is a direct reflection of the underlying state and it is the *presence* of the form that is the product of a variable *insertion* process.[1] Some well-studied examples are final t/d-deletion (or insertion) in American English, syllable-final s-deletion (or insertion) in Caribbean Spanish varieties, and l-deletion (or insertion) in articles and pronouns in Montreal French. There are numerous syntactic examples as well, such as various types of complementizer deletion and "equi" deletion.

The problem of insertion versus deletion is, formally speaking, a special case of finding the best *rule schema* for generating a set of n variants of a linguistic variable. A schema is an ordered set of variable rules, for example,

$$\begin{array}{ll} 1. & u \rightarrow a \\ 2. & a \rightarrow b \\ 3. & a \rightarrow \begin{matrix} c \\ d \end{matrix} \\ 4. & u \rightarrow e \\ 5. & b \rightarrow f \end{array} \qquad (1)$$

New Ways Of Analyzing Sound Change

Copyright © 1991 by Academic Press, Inc.
All rights of reproduction in any form reserved.

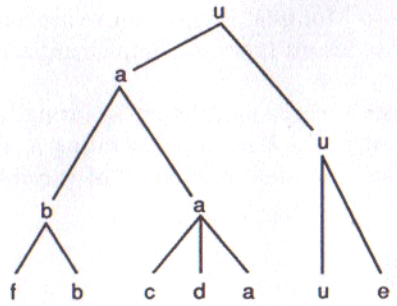
where u is the "underlying" form, and each of the other variants occurs on the right-hand side of exactly one rule and always before it occurs (if ever) on the left.

This schema is interpreted as follows: for each occurrence of the linguistic variable in question, there is the initial choice of whether the underlying form will be converted by rule 1 into a. Whenever rule 1 has successfully applied, some of the a's may then be converted into b's by rule 2. For those a's that remain (i.e., where rule 1 has applied but rule 2 has not) there are then the simultaneous possibilities of changing into c or d, as expressed by rule 3. Finally, cases of b and remaining cases of u may be converted into occurrences of f and e, respectively.[2] Thus, a rule order schema defines, and is defined by, a precise derivational history for each of the *n* variants of the variable.

In this paper we first give a complete characterization of the mathematical structure of such rule order schemata, followed by an account of their implications for the study of natural speech data. We then extend our discussion to the more difficult case where more than one rule can give rise to the same variant. For maximum clarity and generality we proceed at first in an entirely formal manner, and introduce examples only after the exposition of the theory in the simple case (i.e., only one rule creating each variant).

2.1. STRUCTURE OF A SCHEMA

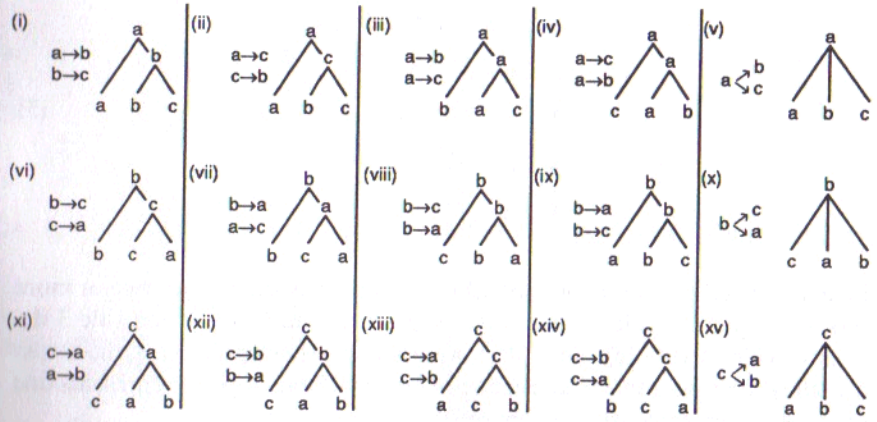
The schema in (1) may also be represented as a tree[3]:



(2)

For a seven-variant variable, there are a total of almost two million different schemata such as (1) to choose among [counting all those in Note 2 as the same as (1), and all those mentioned in Note 3 as the same as (1)]. Even a

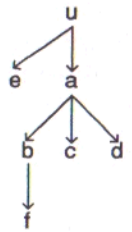
three-variant variable may be accounted for by 15 different schemata:



(3)

In Sankoff & Rousseau (1981, 1982, 1989) we have developed a statistical methodology for choosing the 'best' schema to account for a given set of *n*-variant data.

A fact fundamental to this methodology is that any rule order schema such as (1) consists of two logically distinct hypotheses. One is the *depth hypothesis*: which variants "underlie" which other variants. For (1), this may be represented as



(4)

That is, b is deeper than f; a is deeper than c, d, and b (and f as well, by transitivity); and u underlies all the other variants.[4]

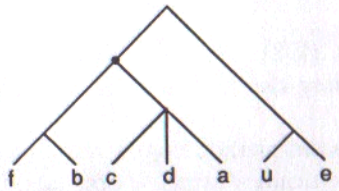
The second hypothesis implicit in any schema such as (1) is its *token classification*: which variants have largely similar derivations, which have

very different derivations, and so on. For (1), the token classification is

1. {u, e}, {a, b, c, d, f}
2. {a, c, d}, {b, f}
3. {a}, {c}, {d}
4. {u}, {e}
5. {b}, {f}

(5)

Each rule in (1) defines a corresponding division of variants into two or more sets. For example, rule 2 defines two sets {a, c, d} and {b, f}, and rule 3 defines three sets {a}, {c}, and {d}.[5] Another way of representing the token classification is as a tree where each nonterminal branch point represents one of the divisions into two or more sets:

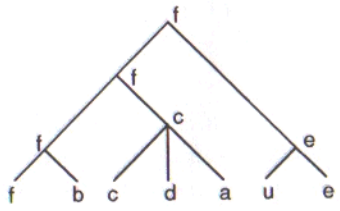


(6)

For example, the node indicated with the dot in (6) represents the division {a, c, d}, {b, f}.

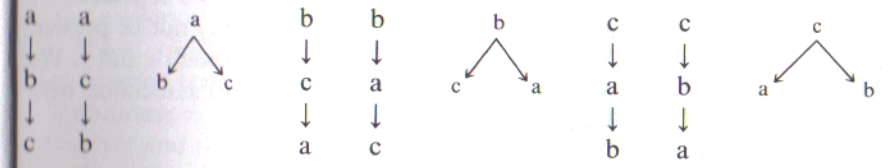
Again, although (6) appears similar to (1), and is indeed implied by it, they are not equivalent. Thus, (6) is also the token classification tree for other schemata, such as

f → e
e → u
f → c
c → d, a
f → b



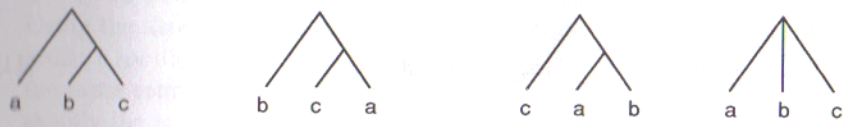
(7)

As another example, consider the 15 schemata in (3). They involve just nine depth hypotheses:



(8)

and only four token classifications:



(9)

2.2. ANALYSIS OF DATA ACCORDING TO A SCHEMA

The statistical analysis of linguistic performance to infer the correct schema is based on the number of tokens of each of the n variants in a number (possibly a very large number) of different contexts. The analysis requires one simplifying assumption that may or may not be reasonable for a given variable, but which we will not discuss further here, namely, that each context is fixed during the application (or nonapplication) of the rules in the entire schema. That is, no changes affect the components of the context of a variable while it evolves into its surface form.

Given that the variants of a variable are generated by some rule order schema in this way, a basic fact, not difficult to prove mathematically, is that the statistical data bear on the token classification but not on the depth hypothesis. That is, statistical analysis can tell us something about rule order, namely, the token classification it implies, but it can tell us nothing about which variants underlie which other ones.

This fact is a consequence of the inferential framework within which we assess the effects of context on the probabilities that a variable will be expressed as each of the n different variants.

We first review the simplest, two-variant case. Suppose we represent by A, B, C, ..., various features or components that may or may not be present in the context of a given instance of a variable with two variants u and v. We assume there is a fixed number (positive, negative, or zero) associated with each of these features, denoted by α, β, γ, ..., such that

$$\log \frac{p}{1-p} = \alpha + \cdots \tag{10}$$

where p is the probability that in a context consisting of features A, ..., variant u will occur rather than variant v, and $1 - p$ is the probability that v will occur and not u. In a context containing B, C, ..., but not A, the formula will be

$$\log \frac{p}{1-p} = \beta + \gamma + \cdots \tag{11}$$

That is, the number associated with a feature contributes to the variant choice probabilities for a context if and only if this feature forms part of this context. Thus, for different contexts we have different probability formulas. We say that the set of probabilities satisfying formulas like (10) and (11) follows the logistic model.

Now, until we have analyzed some data, we do not know the value of p, in the different contexts, nor of the parameters α, β, etc. But if we have a certain number of occurrences of a variable in a given context and r of them are u's and s of them are v's, then $r/(r + s)$ is the best estimate of p, though generally it will not be exactly equal to p because of statistical fluctuation.

If we have quantities of data for *many* different contexts, however, then simply dividing the number of u's by the total number of u's and v's in each context is no longer the best way to estimate the probabilities in this context. This is because the simple method does not take account of the strong relationship between the probabilities in different contexts, relationships that are implicit in the logistic model. As α is a fixed number, variant probabilities in all the contexts containing A will be affected in the same way by the size of α. Indeed it is really α, β, and the other feature effects that we would like to estimate because there are generally a small number of them and they determine precisely the values of the entire set of context probabilities, which may be large and unwieldy, containing all the different possible ways of combining the various context features: phonological, syntactic, lexical, stylistic, sociological, and so forth.

The entire set of data on variant occurrences in all contexts is used to estimate the feature effects using a multiple regression method under the maximum likelihood criterion. Though this is much more complex than just dividing r by $r + s$, it consists of standard statistical procedures that we will not elaborate here.

Two points are crucial, however, for the ensuing discussion. First, the estimation procedure is based, as in the simple case of r occurrences of u and s occurrences of v in a single context, on the number of occurrences of one variant and the other *compared systematically across the set of all contexts.*

Second, when the procedure produces the parameter estimates, it also produces a value for the *likelihood function*, which is a measure of how well the linguistic model—a set of equations like (10) and (11)—fits the data.

We return now to the problem of underlying variants. If we make the hypothesis that u underlies v, and use a given data set to estimate the effects, say $\hat{\alpha} = 2$, $\hat{\beta} = 1$, $\hat{\gamma} = -0.5$, ..., of context features A, B, C, ..., on the rule u → v, we obtain some number, say 0.35, for the likelihood of the model. Using the identical data set to estimate the model parameter under the opposite hypothesis, namely, that v underlies u and v → u, we necessarily obtain the same estimates but of opposite sign: $\alpha = -2$, $\beta = -1$, $\gamma = 0.5$, ..., and exactly the same likelihood (0.35 in this example). This means both underlying hypotheses fit the data equally well (or equally badly) and hence the data do not help us choose which hypothesis is better.

We are now in a position to understand the general n-variant case. For each occurrence of one of these variants, we know exactly which of the rules in a rule schema have applied, which have not, and which were not even eligible in generating this occurrence. Thus in (1), for any occurrence of the variant b, we know that rule 1 (u → a) and rule 2 (a → b) have applied, and rule 5 (b → f) has not. Once rule 1 has applied, there is no question of the application or not of rule 4 (u → e), as the variable has already been transformed from u to a. Similarly, for rule 3 $\left(\text{a} \begin{matrix} \nearrow \text{c} \\ \searrow \text{d} \end{matrix}\right)$, once rule 2 has applied, the variable has become b and is no longer eligible for rule 3.

Thus, for any set of data, we know, for each rule in a schema, exactly how many occurrences were eligible for the rule in each context and how many times it actually applied. We can therefore apply our statistical procedure to estimate the effects of the contextual features on the operation of the rule.[6]

Again, this estimation is qualified by the value of the likelihood. It is a basic mathematical property of the likelihood function that the likelihood that an entire schema (including specified feature effects for each rule in each context) has generated a given set of data is simply the product of the likelihoods of all its component rules.

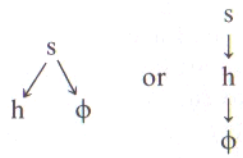
The key to the next part of our argument is illustrated in the comparison of the second highest branch point in (2) and the corresponding dot in (6). In (2), when we say we know which occurrences of a variable were eligible for the rule a → b and which of these represent actual applications of the rule, this means we know that all observed tokens of a, b, c, d, and f represent occurrences that were eligible, and of these, the tokens of b and f represent applications and the tokens of a, c, and d nonapplications. But we have just explained how the likelihood of the rule does not depend on its direction; were a, c, and d the applications and b and f the nonapplications, the likelihood would have been the same. All that matters is the division of the variants into two (or more) subsets represented by the node. The same holds for each node in the tree, and hence the likelihood of the entire schema, the product of the individual likelihoods, depends only on its token classification. Though we cannot make any inference about underlying forms, we can still choose among different rule order schemata by comparing the likelihoods of their token classifications.

2.3. SOME EXAMPLES

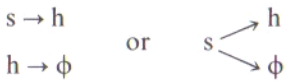
Within this statistical framework a series of studies have been carried out, on syllable-final consonant weakening in Panamanian Spanish (PS) using Cedergren's (1973) data and in Puerto Rican Spanish (PRS) using Poplack's (1979) data, on English complementizers (by Kroch, 1981), and on sequencing conjunctions in Ontarian French (Mougeon, Beniak, and Valois, 1983).

To illustrate the type of information to be obtained in such studies we briefly summarize the results.

1. In PS, under the depth hypotheses

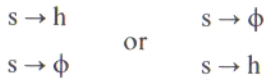


the most likely schemata for s-reduction are the intrinsic order, aspiration followed by deletion, and the unordered analysis



(12)

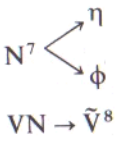
while the extrinsic orders



(13)

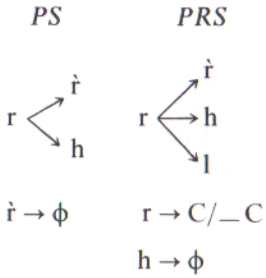
are very unlikely.

2. In PRS, the processes of verb-final *n*-reduction are not linearly ordered, velarization and deletion being simultaneous choices, with only the remaining occurrences being susceptible to vocalization:



(14)

3. In both PS and PRS, r-reduction seems to involve independent alternative processes of spirantization and aspiration, with deletion a later option. In PRS, assimilation of unreduced occurrences is another option and lambdacization is an entirely separate choice concurrent with aspiration and spirantization:



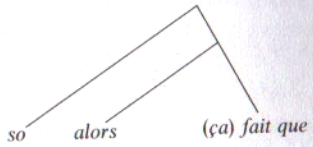
(15)

The different position of deletion in these two schemata is probably due to differences in the coding procedure used for the two sets of data.

4. For English sentences where *which*, *that*, and a null complementizer may alternate, performance evidence provides no clear evidence that any pair of these are derived more similarly than the other. None of the four possible token classifications is much more likely or much less likely than the other. This may be interpreted in terms of the relative force of surface constraints versus derivational history on the distribution of these forms.

5. Of the three conjunctions, *alors*, (*ça*) *fait que*, and *so*, the distribution of the first two is quantitatively more similar than that of the English

loan-word, so that the most likely token classification is

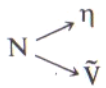


(16)

2.4. MIXED RULES

In the discussion to this point, a basic assumption has been that the rule order schema can always be represented by a tree as in (2) [or as in (4) together with (6)].

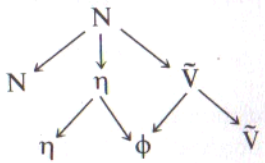
One can very well imagine, however, that some sets of rules may not satisfy this assumption. For example, instead of (14), n-deletion in PRS might conceivably be the end result of both velarization *and* vocalization:



$$\eta \rightarrow \phi$$
$$\tilde{V} \rightarrow \phi$$

(17)

This would not be represented in terms of a tree, but only in terms of a more general type of structure, a network:



(18)

Analyzing variation data according to a network is much more difficult than in the case of a tree. The presence in the schema of two or more rules that give the same output means that we can no longer know for each token exactly which sequence of rule applications and nonapplications generated it. In the network in (18), we know how tokens of N, η, and $\tilde{V}$ were generated, but tokens of ϕ could have arisen either through $\eta \rightarrow \phi$ or through $\tilde{V} \rightarrow \phi$. Thus, for a given data set with observed numbers of N, η, $\tilde{V}$, and ϕ, we do

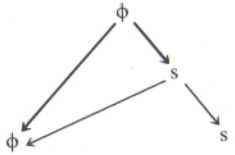
not know how many η were converted to ϕ and how many $\tilde{V}$ were converted to ϕ; hence we do not even know how many η and how many $\tilde{V}$ were produced by the first rule. We therefore cannot statistically analyze any of the rules, nor, of course, calculate the likelihood of the schema, using the same methods as before.

There is, however, a way of inferring much of the apparently missing information in such cases, without any additional data. This is based on an important assumption inherent in the logistic model, namely, that the effect of any contextual feature A on the variant choice probabilities is determined entirely by the presence of the associated parameter α when A is present in the context and by the absence of α when A is absent. The effect α is neither augmented nor diminished by the presence of any other feature, though each of these may make their own independent contribution to *p*. This assumption has been intensively scrutinized in the course of a decade of sociolinguistic data analysis, and has generally held up with respect to features in the *linguistic* context, especially when the set of all contexts has been most appropriately characterized from both the linguistic and the statistical points of view. Independence of effects also obtains among linguistic features, stylistic factors, and the tendencies of individual speakers. Where the independence hypothesis has not held up in a number of studies is among the demographic factors characterizing the speaker in the sample of the speech community. Methods of studying the interaction among age, sex, social class, and other characteristics are available, as are techniques for removing the statistical effects of this interaction. The net result is that the judicious use of the independence assumption as a general working hypothesis is as justified with linguistic data as it is in statistical analysis more generally.

What are the consequences of this assumption for the analysis of "mixed" rules, such as the one(s) producing ϕ in (17)? The answer is illustrated by the following fact: suppose the rule $\eta \rightarrow \phi$ operates according to a logistic model in a given set of contexts and the rule $\tilde{V} \rightarrow \phi$ operates according to another logistic model (i.e., different parameter values). Further, suppose we knew the derivational history of every token, even the ϕ's. We could divide the ϕ's into ϕ_η's produced by the rule $\eta \rightarrow \phi_\eta$ and ϕ_V's produced by the rule $\tilde{V} \rightarrow \phi_V$. Then the token partition $\{\eta\}, \{\phi_\eta\}$ would be well fit by a statistical analysis according to the logistical model, and the partition $\{\tilde{V}\}, \{\phi_V\}$, by another logistic model. But the token partition $\{\eta, \tilde{V}\}, \{\phi\}$ (i.e., $\{\eta, \tilde{V}\}, \{\phi_\eta, \phi_V\}$) would *not* in general be well fit by a logistic model, basically because there would be a systematic breakdown of the independence assumption. Even though this assumption may be valid for $\eta \rightarrow \phi_\eta$ and $\tilde{V} \rightarrow \phi_V$ separately, incorrectly treating the two rules as one introduces an apparent interaction between the feature effects. The nature of this artificial interaction can be statistically predicted in terms of the parameters of the $\eta \rightarrow \phi_\eta$ and $\tilde{V} \rightarrow \phi_V$ rules.

This means that with real data, generated by the schema in (17), where we cannot distinguish between the ϕ_η and ϕ_V, and where we therefore do not know the derivational history of each token, an analysis of the partition $\{\eta, \tilde{V}\}, \{\phi\}$ should yield poor results, in that the model will not in general fit the data very well. Moreover, the disagreement between data and model will take the form of feature interaction effects of the type predicted in the imaginary situation where the ϕ_η and ϕ_V are distinguishable. The detection of such systematic interaction may then be used as an indicator or test for the existence of two rules rather than just one generating ϕ.

We have been discussing the schema in (17) and (18) to illustrate the notions of mixed rules and networks. The mathematical details and statistical analyses have been completely worked out in a somewhat simpler case, which is, however, of considerable linguistic interest. This is the case of plural s marking in Caribbean Spanish. We might imagine that variable presence of s in this case is due both to the variable application of the (syntactic) marking and agreement rules and to the variable application of the (phonological) deletion rule which applies to all syllable-final s's. This is depicted in



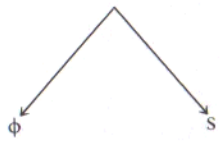
(19)

where the heavy lines indicate the output of the syntactic rule, and the light lines, the output of the phonological rule.

To test this possibility against a single-rule phonological[9] solution (i.e., where syntactic marking is assumed obligatory) on data from PS, we compare the network in

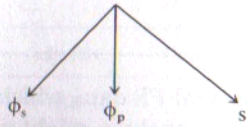
(20)

with the tree in



(21)

The network in (20) results from the indistinguishability in



(22)

of the ϕ_s tokens, representing nonapplication of plural marking, and the ϕ_p tokens, representing the phonological deletion of the plural marker.

To illustrate the logic of the comparison we will use, we first construct two hypothetical data sets, one based on (20) and one on (21).

Suppose first that (20) and (22) represent reality. In (23) we list the three variants in the left-hand column, with the effects[10] assigned to each of the contextual features arranged beside it. Note that we have invented these figures, though they are quite plausible given what is known or may be assumed about this variable. Thus ϕ_s, representing nonapplication of plural making, is not differentially affected by different phonological contexts; ϕ_p, representing deleted tokens, is similarly insensitive to syntactic context.

(23)

	Carrier of Marker (Syntactic Context)			Following Phonological Context		
	Determiner	Noun	Adjective	Consonant	Vowel	Pause
s	.56	.35	.14	.14	.56	.35
ϕ_s	.14	.35	.56	.30	.30	.30
ϕ_p	.30	.30	.30	.56	.14	.35

On the basis of these figures, and using standard methods of statistical simulation, we then generate an artificial data set[11] containing the same number of tokens in each context as Cedergren's (1973) PS data. The breakdown in terms of the three variants is determined by an independent trinomial experiment in each context, whose probabilities are determined by the figures in (23).

We then combine the tokens of ϕ_s and ϕ_p to produce a data set that perfectly typifies the network situation in (20). In other words, we now have an artificial data set that we know was produced according to the network hypothesis but in which we cannot distinguish the ϕ_s and ϕ_p tokens.

Applying an ordinary two-variant analysis to the token classification {s}, {ϕ}, we obtain the estimates in (24), which are quite reasonable. This could

	Determiner	Noun	Adjective	Consonant	Vowel	Pause
s	.74	.54	.23	.25	.73	.52
ϕ	.26	.46	.77	.75	.27	.48

(24)

bolster the suspicion that the real PS data, which give results very similar to (24) when analyzed by the same analysis, might also be generated by the network in (20). How can we detect the twofold origin of the ϕ tokens?

In the equations in (10′) of note 6, let p, q, and r represent the probabilities of s, ϕ_p, and ϕ_s, respectively, used in generating the artificial data set. We can compare (10) with these equations as follows:

$$\begin{aligned}\log\frac{p}{1-p} &= \log\left(\frac{p}{q}\cdot\frac{q}{1-p}\right)\\ &= \log\frac{p}{q}+\log\frac{q}{1-p}\\ &= \alpha^{(1)}+\beta^{(1)}+\cdots+\log\frac{q}{1-p}\end{aligned} \tag{25}$$

Thus, whereas (10) relates p to a simple sum of parameter effects, (10′) or (25) relates it to a similar sum plus the term $\log(q/(1-p))$, which may be termed an interaction effect. This effect in turn must satisfy the second equation in (10′):

$$\begin{aligned}\log\frac{q/(1-p)}{1-q/(1-p)} &= \log\frac{q}{1-p-q}\\ &= \log\frac{q}{r}\\ &= \alpha^{(2)}+\beta^{(2)}+\cdots\end{aligned} \tag{26}$$

We can estimate the parameters $\alpha^{(1)}$, $\beta^{(1)}$,..., as well as $\alpha^{(2)}$, $\beta^{(2)}$,..., in (10′) using only the data on occurrences of s and ϕ because we now know from (25) and (26) just how the probabilities p (of s) and $1-p$ (of ϕ_s and ϕ_p combined) depend on these two sets of parameters. As there are more parameters in this model than in (10), the fit of data to model is guaranteed to be better, with a higher likelihood. But we can test if this increase in likelihood is statistically significant, taking account of the extra parameters: if there are m independent parameters in the model represented by (10), then there are $2m$ in the model represented by (10′), and twice the increase in the logarithm of the likelihood can be tested by a chi-square test with m degrees of freedom.

When this analysis is applied to the invented data, we recover the parameter effects

	Determiner	Noun	Adjective	Consonant	Vowel	Pause
s	.56	.36	.14	.13	.58	.35
ϕ_s	.15	.35	.54	.28	.31	.30
ϕ_p	.24	.30	.32	.59	.13	.35

(27)

which are strikingly similar to those in (24) and are a confirmation of our ability to do a three-variant analysis, even if the distinction between two of the variants is missing. More important is that the likelihood of the analysis associated with (27) is significantly greater than that associated with (24), proving what we already know, that the invented data were generated by the network in (20) and not by the tree in (21).

Now that the methodology has been demonstrated to be capable of detecting mixed rules for artificial data, what of the real data?

Applying exactly the same procedure for the PS data, we obtain the feature effects

	Determiner	Noun	Adjective	Consonant	Vowel	Pause
s	.36	.22	.23	.32	.40	.26
ϕ_s	.50	.07	.57	.33	.19	.51
ϕ_p	.14	.71	.19	.35	.40	.23

(28)

for the analysis according to (25) and (26). Not only is this linguistically counterintuitive, but a chi-square test shows that it does not fit significantly better than an analysis according to (10), which results in values similar to those in (24). Thus we can reject the network hypothesis for these data. s-Deletion in PS seems to result from a single (phonological) process, rather than from the combined effects of variable syntactic and phonological rules.

2.5. CONCLUSIONS

The simple example we have analyzed is a striking illustration of how statistical methods can distinguish between two qualitatively different hypotheses on the organization of a grammar.

In generalizing the notion of rule order schema to permit "mixed rules," we open up an entire new problem area for the methodology of multivariant

linguistic variables. The mathematics in (25) and (26) is not sufficient to analyze more complicated models, such as the one for N-deletion we discussed earlier. A systematic investigation of the equations for various mixed rule configurations, even for only three or four variants, promises to be a major undertaking. Even the combinatorial problem of enumerating the different logically possible configurations is not trivial.

NOTES

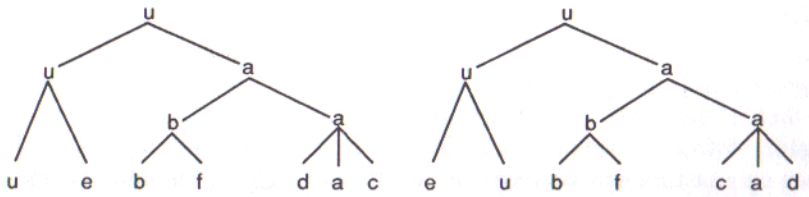
[1] For insertion to be a plausible explanation of the data requires the morphological, orthographic, or simply normative recoverability of the range of contexts that are eligible for insertion. Indeed, the occurrence of hypercorrect forms is often cited as evidence for the insertion argument, in communities where the vernacular is far removed from normative varieties.

[2] Rules 1, 2, and 3 must occur in this relative order, but rule 4 can occur anywhere after rule 1, rule 5 can occur anywhere after rule 2, and c and d can be interchanged in rule 3, without changing the interpretation of the schema. Once an occurrence of the variable has been converted into an a by rule 1, further rules rewriting u and those rewriting a cannot interfere with each other, so their order is of no importance. We thus treat the following schemata as identical to (1):

u → a	u → a	u → a	u → a	u → a
u → e	a → b	a → b	a → b	u → e
a → b	u → e	b → f	b → f	a → b
b → f	b → f	u → e	a → d, a → c	a → c, a → d
a → c, a → d	a → d, a → c	a → c, a → d	u → e	b → f

etc.

[3] Note that the branches of the tree represent the derivational steps in the production of surface forms through (1) or, for that matter, through any of the schemata in Note 2. The left-to-right order of (2) on the page is immaterial, however, so that (1) could be equally well represented by, for example,



, etc.

[4] Although each rule order schema implies a specific depth hypothesis, the two are not equivalent. For example, many other rule order schemata *different* from (1) have the *same* depth hypothesis (4), for example,

u → e	u → a	u → a
u → a	a → b	a → b
a → c, a → d, b	a → c	a → d
a → c, a → d, d	a → d	a → c
b → f	b → f	b → f
b → f	u → e	u → e

etc.

[5] The order of the elements within each set is of no consequence; neither is the order of the sets. Thus, {a, c, d}, {b, f} is equivalent to {a, c, d}, {f, b} and to {b, f}, {a, d, c}.

[6] The statistical procedures are somewhat more complicated for rules like rule 3 involving more than two possible outputs. For example, if there are two output variants as in rule 3, there are not one but two sets of feature effects $\alpha^{(1)}, \beta^{(1)}, \gamma^{(1)}, \ldots$, and $\alpha^{(2)}, \beta^{(2)}, \gamma^{(2)}, \ldots$. In addition, instead of two probabilities, p and $1 - p$ of application and nonapplication, there are three probabilities, say p, q, and r, representing the three subsets of tokens created by the rule. Of course, $p + q + r = 1$. Instead of an equation like (10) we have two equations:

$$\log(p/q) = \alpha^{(1)} + \beta^{(1)} + \cdots, \qquad \log(q/r) = \alpha^{(2)} + \beta^{(2)} + \cdots \tag{10'}$$

For a rule with three output variants, there would be three sets of parameters, four probabilities, four subsets of tokens, three equations, and so on. Nevertheless, the estimation procedures are analogous to the single-output case, and a single likelihood measures how well the rule accounts for the pertinent data.

[7] N represents the hypothesized underlying alveolar segment that is obligatorily assimilated to the following consonant.

[8] For simplicity, this rule will henceforth be written $N \to \tilde{V}$. Also $\tilde{V} \to V$ will be written $\tilde{V} \to \phi$ as we are focusing only on what happens to the nasal segment, not on the vowel.

[9] Of course, this phonological rule may be syntactically *conditioned*: the grammatical category of the plural word will have an effect on the deletion probability.

[10] Rather than give the values of $\alpha, \beta, \ldots$, as in (10) or (11), it is common practice to present $\log[\alpha/(1-\alpha)], \log[\beta/(1-\beta)], \ldots$, as the latter are between 0 and 1 and have the form of probabilities.

[11] For these data, and for the analyses we discuss later, we postulate and calculate a social class effect as well as the linguistic effect, but because this is not directly pertinent to our topic, we shall omit the details.

REFERENCES

Cedergren, H. J. (1973). *Interplay of Social and Linguistic Factors in Panama.* Ph.D. thesis, Cornell University.

Kroch, A. (1981). Paper presented at the Xth NWAVE Conference, Philadelphia.

Mougeon, R., Beniak, E., and Valois, D. (1983). Paper presented to the Canadian Linguistic Society, Vancouver.

Poplack, S. (1979). *Process and Function in a Variable Phonology.* Ph.D. thesis, University of Pennsylvania.

Sankoff, D., and Rousseau, P. (1982). Un diagnostica para reglas convergentes. In *El Español del Caribe* (O. Alba, ed.), pp. 291–300. Universidad Madre y Maestra, Dominican Republic.

Sankoff, D., and Rousseau, P. (1981). Binomial decomposition for multinomial data in logistic linear models. In *Topics in Applied Statistics* Y. P. Chaubey and T. D. Dwivedi (eds.) pp. 199–208. Montreal.

Sankoff, D., and Rousseau, P. (1989). Statistical evidence for rule ordering. *Language Variation and Change* **1**(1), 1–18.