Advanced Varbrul Analysis

Gregory R. Guy
Stanford University

0. Introduction. This chapter is based on a workshop given at NWAV-XVI. It is intended as a practical, user-oriented survey of some of the methods and problems associated with doing quantitative 'variable rule' analyses of linguistic data using the Varbrul program developed by Sankoff and Rousseau. The material covered here should not be considered as remarkable original discoveries by the author; on the contrary, most of these topics form part of the standard technical lore of statistical analysis. But among linguists this information has primarily been transmitted by word-of-mouth, an academic version of the preliterate tradition of oral history. This is a hindrance to the rapid dissemination of an essential technique to a community of interested scholars. The information needs to be available in a written version accessible to linguists. That is what this chapter tries to do. I write this not as a trailblazer or innovator, but rather as the humble recorder of the tales told by the people who know. Any faults of interpretation are of course my own.

The term 'advanced' in the title may require clarification. The present work is intended for the person who has read about and perhaps used Varbrul methodology, and would like to know more about how to make it work. It is not so advanced that an experienced researcher in the field will necessarily benefit from it, but does go beyond the level of mere introduction to the subject. Thus I assume at least a minimal acquaintance with the goals, techniques, and terminology of Varbrul analysis. A very brief review of this material is provided below, but a reader who lacks basic familiarity is advised to review the literature on the subject, such as Cedergren 1973, Cedergren & Sankoff 1974, Guy 1975, Sankoff 1975, Rousseau and Sankoff 1978, Sankoff & Labov 1979, Rousseau 1983.

1. Why 'variable rules'? Variable rule analysis was developed in linguistics as a way of accounting for structured, rule-governed variation in language use. This is linguistic variation that regularly shows greater or lesser rates of occurrence in particular environments, or that regularly predominates among particular social groups or in particular speech styles.

An example is the well-known case of postvocalic /r/ in New York City. This originally 'r-less' dialect is beginning to import 'r-ful' pronunciations as a prestige norm from the surrounding dialects sometimes called General American. The use of consonantal realizations in New York is finely stratified by social class, adjusted in each situation according to speech style, and conditioned by linguistic context such as stress and following segment. These patterns are highly systematic across all social groups, as is illustrated in Figure 1, from Labov 1966.

This is not 'random' or 'free variation', but is rather what Weinreich, Labov and Herzog term 'orderly heterogeneity' (1968:100). There are clearly some 'linguistically [and socially] significant generalizations' to be captured here. Higher class groups use more /r/ all the time, and everyone uses more /r/ in their more formal styles. A descriptively adequate account of this variation should precisely define the nature and extent of each of these conditioning effects, and allow us to probabilistically 'predict' the approximate rate of use of /r/ by an individual, given certain information about the social characteristics of the person, the social situation, and the linguistic environment.

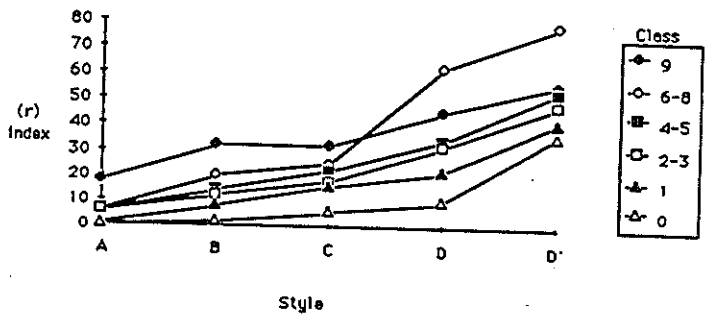

Figure 1. Class stratification of (r) in New York City.
(after Labov 1966:240)

The goal of variable rule analysis is to do precisely this. Dealing with variable, and not categorical phenomena, it is necessarily quantitative. Any generalization we might formulate about Figure 1 could not be disproved by a single counterexample, but only by a statistical test of a body of data. Hence the analysis will involve counting instances of the variable and statistically describing the range of the variability and the constraints or 'factors' that influence it.

An analytical problem arises here, from the fact that we cannot do controlled experiments that isolate a single factor and test its effect on the variable in question. Every token of a variable occurs embedded in an utterance and a social context which could encompass a large number of factors influencing the speaker's selection from the range of alternants. Therefore the analysis must necessarily be MULTIVARIATE; in other words it is an attempt to model the data as a function of several simultaneous, intersecting, independent forces, which may be pulling in different directions. In fact, one product of the analysis is a numerical measure of the strength and 'direction' (favoring or disfavoring) of each force.

Such an analysis must be based on a large corpus of observations of the variable in many different contexts, and involves partialling out the several effects by controlling for each in turn. How this is done can be illustrated by re-examining Figure 1. We discern the effects of style by tracing along the line for each class group in turn. Going from left to right along any line, the slope is always upward. Thus we conclude that, regardless of a speaker's social class, more formal styles always evoke more /r/-pronouncing. A counterexample would be a line sloping downward; in fact there are none.

Similarly we identify the class effect by controlling for style. Looking up the column for style B, for example, we find the different socioeconomic groups neatly stratified from highest to lowest status. When we note that the same pattern is also true in styles C and D, we begin to generalize that status is correlated with /r/ pronouncing. In this case however there are slight discrepancies at the stylistic extremes. Whether these are significant

violations of the pattern or not is a question that can be answered by statistical tests.

The important point here is that only these controlled comparisons allow us to make valid generalizations. Comparing class 9 speakers in style B with class 1 speakers in style D would not tell us anything about style or class. Rather, one or the other dimension must be held constant.

For some readers it may be useful to review the terminology of variable rule analysis, introduced by Labov 1969 and Cedergren & Sankoff 1974. A 'variable rule' is a context-sensitive rewrite rule relating a pair of variants such as x —> <y>, so that when the rule applies 'y' appears, and when it fails to apply, 'x' appears. Each of the analytical dimensions in the context which affects whether or not the rule applies is called a 'factor group'. In the example, STYLE and SOCIOECONOMIC CLASS are the two factor groups. Each factor group is made up of a set of discrete 'factors'; e.g. the style factor group comprises five factors, A, B, C, D, and D'. Each unique combination of factors describes a 'cell'; e.g. [SEC 9, style A] is one cell from Figure 1. In the quantitative analysis each factor is assigned a 'factor value' (or 'probability')—a number between zero and one characterizing the effect of that factor on the variable rule in question (the higher the number, the more likely the rule is to apply when that factor is present in the context). And finally, an entire data set is characterized by an 'input probability' ('p_0'), which is in effect a global measure of rate of rule application. A mathematical function is used to combine the factor values and input probability to yield expected rates of rule application in individual cells. Several such functions have been used in the literature; the one currently favored is the 'logistic function' (see Rousseau and Sankoff 1978 for the mathematical details).

In the balance of this paper I will discuss five issues that arise in the course of doing this kind of analysis. Three of them stem from the analytical need for controlled comparison: orthogonality of factor groups, skewing of data distribution, and over-lapping data distribution between the input probability and some other factor. The fourth is a problem in the statistical estimation of constraint effects: deciding what is significant. And the last is a universal problem of multivariate statistics: independence versus interaction.

2. Orthogonality of factor groups. To do the kind of controlled comparisons we have described above, the factor groups must be 'orthogonal', or nearly so. That is, they must freely co-occur, and not be sub- or super-categories of each other. In the example, a speaker in any SEC can use any speech style; all possible combinations of SEC and Style occur in the data. Style is not a subtype of SEC, and a given style is not used exclusively by just one or two SEC groups. This is true orthogonality, and it allows the maximum scope for the kind of controlled comparisons we described above. But it often happens that this ideal is not achieved. Then we must confront the question of how far we can deviate from this standard and still obtain valid results.

One way of visualizing this question is to actually arrange two factor groups orthogonally; that is, make them into a two-dimensional table with one factor group on each dimension. If we did this with the data on which Figure 1 is based, all cells in the table would be filled, because data exist for all combinations of Style and SEC. But in a data set with some empty cells, problems with non-orthogonality might arise.

The worst case would be something like the illustration in Table 1. Here each factor in factor group 1 co-occurs with only one factor in factor group 2. In other words, all tokens of 'a' are also tokens of 'x', and all tokens of 'x' are also tokens of 'a'. The same relationship exists between 'b' and 'y'. This presents an impossible analytic problem. For assigning effects to 'a' and 'b', we need to control for 'x' and 'y'. We need to be able to compare what 'a' does when it is associated with 'x' with what 'b' does when associated with 'x', and so on for all other possible controlled comparisons. In this data set, such

comparisons cannot be made; 'a' and 'x' always occur together, and 'b' and 'y', and their effects can never be separated.

Table 1. Non-orthogonal factor groups: 'worst case'.

		Factor Group 2	
		x	y
Factor Group 1	a	√	(no data)
	b	(no data)	√

For the purposes of the Varbrul program, it doesn't matter whether such non-orthogonalities occur by chance (because you simply happened not to collect any tokens in the context of [ay] or [bx]), or by structural impossibility (because [ay] and [bx] represent logical impossibilities, or combinations that do not occur in your language.) Whatever the origins of the problem, no analytical procedure can partial out separate effects for two factor groups that show this kind of data distribution.

It should be noted, however, that a few empty cells in a table do not always create a problem of orthogonality. Table 1 would be analyzable if, say, cell [ay] had data in it, even if [bx] remained empty. Controlled comparisons would still be possible for 'x' and 'y' (in the environment of 'a'), and for 'a' and 'b' (in the environment of 'y'). While we would have less confidence in such an analysis than in one that had all cells filled, we might still obtain useful results.

These issues can be made more concrete by considering examples from real data. One type of non-orthogonal pattern is illustrated in Table 2, from Poplack 1980. Here we have two factor groups that may affect the deletion of the plural -s suffix in Puerto Rican Spanish: position of the token in an NP, and presence or absence of other plural markers preceding the token in the NP. Obviously most combinations of factors in this table are impossible. A token in first position (e.g. las in las casas blancas) cannot have any plural markers preceding it, and a token in second position can have only one marker preceding, either present or deleted. So in fact, only six of the 18 cells in the table have data in them, as these are the only possible combinations.

Table 2. Puerto Rican Spanish: plural -s deletion. Effect of position in NP and presence of preceding marker (after Poplack 1980).

Position of Token in String:	1	2	3
Marker(s) preceding token:			
'nothing'	.24		
S		.44	
0		.52	
0S, SS			.40
S0			.68
00			.73

These 'factor groups', as I have defined them, are clearly non-orthogonal. They are close to the 'worst case' situation described above. Tokens of 'first position' are always preceded by 'nothing' and tokens preceded by 'nothing' are always 'first position'. Most of the cross-comparisons we would like to make are impossible. We can never see how 'third

position' would behave when the preceding marker was simply 'S', for example. Thus the analytical procedure underlying Varbrul would be defeated in this type of data encoding.

A second example comes from my own work on Brazilian Portuguese. Early on I realized that the variables I was looking at showed a certain systematic sex differentiation (with women favoring the standard variants more than men), but there remained a lot of residual idiosyncratic differences as well. So at one point I thoughtlessly attempted to analyze the data with one factor group for 'informant' and one for 'sex', as illustrated in Table 3. Half the cells are of course empty, in fact impossible. Each informant of course has only one sex, and the analysis can never compare the behavior of, say, informant 6 as a male with the same informant's behavior as a female. Consequently the factor values calculated for the sex factor group were meaningless.

Table 3. Non-orthogonal factor groups: subcategorization.

		Sex	
		M	F
Informant Number	1	√	
	2	√	
	3	√	
	4		√
	5		√
	6		√

Having understood the problem that non-orthogonality presents, we should consider how to find it when it occurs, and how to resolve it. The simplest way to find it is of course to carefully examine any coding scheme one has established before attempting a Varbrul analysis. But if one's self-vigilance fails on some occasion, or if the non-orthogonality arises from an unnoticed chance maldistribution of the data, then the problem will manifest itself in the Varbrul output as (1) meaningless results, and/or (2) non-convergence. When either of these occurs, they should be taken as a warning to check for non-orthogonal factor groups.

Curing this problem is relatively simple. The basic thing that we want to avoid is any situation where a factor in one group co-occurs exclusively with just one factor in some other group. But we should notice that there are several different types of distribution involved. The worst case, complete identity between two factor groups as illustrated in Table 1, involves a one-to-one relationship in both directions: every token of 'a' is an 'x' and every token of 'x' is an 'a', and the same for 'b' and 'y'. When there is complete identity between two entire groups, it is best resolved by simply dropping one of the groups from the analysis.

The second type of problem is sub-categorization, as in Table 3. Basically what happens here is that each sex category has been broken down into the individuals who make it up, so the codes in the informant factor group are all subcategories of exactly one category in the sex factor group. This is a many to one relationship: all the data produced by informant one was also coded as being produced by a male speaker, but not all the data produced by male speakers was produced by informant one. The simplest solution to this kind of problem is again using a single factor group comprised of just the superordinate categories, in this case {M, F}, or just the subordinate categories {1,2,3,4,5,6}, or some combination of the two so long as no superordinate category is used simultaneously with some of its subordinates. (Thus a factor group composed of factors {1,2,3,F} would be perfectly workable, but not {1,2,3,4,F}, because informant 4 could then be coded either 4 or F.)

Finally there are combinations of the two, with each other or with normal orthogonal factors in the same groups. The Spanish example in Table 2 combines the two types of non-orthogonal arrangements (preceding markers 'nothing' and position '1' are identities, while positions '2' and '3' each have several subcategorizations in the 'preceding markers' factor group.

Resolutions of such combinations will depend on the individual case. Again, one possibility is reduction to a single factor group. This is actually how Poplack obtained the factor values shown in Table 2. The analysis included only the 'preceding markers' factor group, with no separate factor group included for 'position'. The results are simply displayed with separate columns for position to illustrate that a positional pattern appears to obtain in the data (later positions favor deletion).

But suppose we felt that there really were two different (i.e. orthogonal) effects involved in a data set like Poplack's, and did not wish to ignore the pattern that is apparent in Table 2. How could we capture this within the framework of controlled comparisons necessary for multivariate analysis? Doing this involves redefining the factor groups. In the Poplack data, for example, we could adopt a 'Position' factor group {positions 1,2,3}, plus a 'Preceding markers' factor group redefined in any of the following ways:

- 'yes': a preceding marker occurs somewhere in the string; 'no': there is no preceding marker anywhere in the string. (This would combine Poplack's categories 'S', '0S', 'SS', and 'S0' as 'yes', and her 'nothing', '0' and '00' categories as 'no.)

-'yes': a plural marker is present on the immediately preceding word; 'no' there is no plural marker on the immediately preceding word. (This combines Poplack's 'S', 'SS' and '0S' categories as 'yes', and 'nothing', '0', '00' and 'S0' as 'no'.)

-'yes': a plural marker is present on the immediately preceding word; 'no' there is no plural marker on the immediately preceding word; 'does not apply' (coded as a blank in most versions of Varbrul): no preceding word exists. (This is the same as the previous coding scheme except for preserving a unique status for the first position tokens. When some factor group is coded 'does not apply' for a particular cell, no calculations are undertaken for that factor group in connection with that cell. This does not violate orthogonality.)

To sum up, one should always attempt to devise coding schemes that avoid one-to-one relationships between factors in different groups. If such distributions occur anyway, resolve them by eliminating a factor group, redefining a factor group so that it is actually orthogonal, or using the 'does not apply' option to replace one member of an offending pair of factors.

3. Skewed data distribution. I have just discussed orthogonality in categorical terms. But being a variationist, I must not fail to emphasize that orthogonality can also be looked at as a scalar or relative property. We have said that it is impossible to analyze a data distribution like the one in Table 1, where the cells [ay] and [bx] have no data in them. But what would happen if they had just a tiny amount of data in them, as in Table 4. Will a Varbrul analysis work here?

Table 4. Skewed data distribution. (Numbers of tokens)

		Factor Group 2	
		x	y
Factor Group 1	a	100	2
	b	2	100

The issue here is how the data are distributed across categories in different factor groups. The ideal analytical case is one with a balanced distribution with all cells having equal N's, illustrated in Table 5. The worst case scenario in Table 1 has a coincident distribution between factors 'a' and 'x', and between factors 'b' and 'y'. In between are various degrees of skewed or overlapping distributions, of which Table 4 is one example, with a 98% overlap between 'a' and 'x', and between 'b' and 'y'. So from this perspective the non-orthogonality problem resolves itself into the question of what degree of overlap can we tolerate and still obtain useful results.

Table 5. Balanced data distribution. (Numbers of tokens)

		Factor Group 2	
		x	y
Factor Group 1	a	50	50
	b	50	50

To illustrate what happens with various degrees of overlap, I prepared a series of hypothetical data sets, each with two factor groups {a,b} and {x,y} and four cells. All data sets had the same cell rates of 'rule application', as shown in Table 6, but they differed in how many tokens were allocated to the [ax] and [by] cells (diagonal 1), as opposed to the [ay] and [bx] cells (diagonal 2). The first set had a balanced distribution: i.e. a 50:50 ratio between the two diagonals. The second had a 10:90 ratio, the third 5:95, the fourth 1:99, and the last had a coincident distribution (i.e. 0:100). All were analyzed with Varbrul2M (a Macintosh version of Varbrul2 under development by the author). The results are summarized in Table 7.

Table 6. Percentages of rule application for test data sets.

		Factor Group 2	
		x	y
Factor Group 1	a	66.6%	50%
	b	50%	33.3%

Table 7. Test data set results, showing effect of increasing skewing of data distribution.

Skew ratio (diag. 1: diag. 2)	Significance step-up	step-down	Factor Values a	b	x	y	Conver-gence (iter. #)	Chi^2 per cell
50:50	yes	yes	.59	.41	.59	.41	4	0
10:90	no	yes	.58	.42	.58	.42	15	.004
5:95	no	yes	.58	.42	.58	.42	no	.012
1:99	no	no	.53	.47	.53	.47	no	.121
0:100	-	-	.50	.50	.50	.50	(1)	(0)

With a balanced distribution we get robust and meaningful analytical results, with factor values demonstrating, as expected, that factors 'a' and 'x' substantially favor the 'rule'. Both factor groups are significant, convergence is achieved promptly, and the 'chi-square per cell' error factor is nil, indicating that the model fits this data set perfectly. But as we start getting highly overlapped distributions, all of these results begin to collapse. At the 90% level, there is a modest effect on the factor values, significance is becoming rubbery, convergence is much harder to achieve, and there is a modest increase in the error factor. At the 95% level convergence is lost entirely and the error factor triples. At the 99% level, significance is also lost entirely, the factor values are greatly altered, and the error factor is again multiplied by 10. Finally, at the 100% level the factor values are meaningless and insignificant, and there is no point in attempting analysis beyond the first iteration.

What do these results imply for the linguist working with real data? The details of what happens will vary from case to case, but we can adopt certain 'rules of thumb'. A 90% overlap is tolerable, although one should realize that some distortion of results is probably occurring, and the analysis is already taking a lot longer. 95% is probably the absolute limit of reasonable analyzability. A researcher would be very ill-advised to rely heavily on (or publish) figures based on a data distribution with more than 95% overlap between two factors in different groups.

4. Data overlap with p_0. One type of overlapping distribution that often goes unnoticed but can have the same distorting effects on analytical results is an overlap between a factor and the 'input probability' - p_0. This will occur whenever there is a factor group with one factor 'i' that encompasses 95% or more of the data, plus one or more other factors ('j','k','l',...) in the group that together comprise less than 5% of all tokens. A coincident distribution arises here because this factor 'i' is found in the environment of 95% of all tokens, while p_0, by definition, is in the environment of 100% of all tokens. So almost all the time, when the algorithm is attempting to partial out effects, it cannot decide whether to assign an effect to 'i' or p_0.

My first encounter with this problem was in connection with an analysis of some Brazilian Portuguese data. I was looking for functional constraints on plural -s occurrence, under the hypothesis that speakers would omit plural markers more freely if the information about plurality was recoverable from elsewhere in the sentence or discourse. Some of the analysis results appear in Table 8.

Table 8. Brazilian Portuguese: Noun phrase agreement. Effect of place of additional plural information. (from Guy 1981.)

	% Plural Marked	N	Prob.
Add'l plural info precedes NP	61	152	.24
No add'l plural info.	61	8588	.63
Add'l plural info follows NP	84	1046	.65

Looking at the factor probability values, we find that plural marking appears to be strongly disfavored by the presence of additional plural information in the discourse preceding the NP that contains the token, but when there is no such information, or when such information follows the NP, plural marking of the token is favored. The factor values for these two favoring categories are very close (.63 and .65); in fact, a significance test reveals that they are not significantly different. This suggests that there is a simple binary effect here: when the plurality of an NP has been clearly established in the discourse before

it is uttered, it won't have to be morphologically indicated very often within the NP itself; otherwise morphological markers will be more important.

However, if we implement such an analysis, a major distribution problem arises. Inspecting the N column reveals that combining the last two factors in the group yields a superfactor that characterizes 98.5% of the data in the whole corpus, arrayed against another microfactor comprising a mere 1.5%. The superfactor is thus nearly coincident with the input probability, and all of the effects of overlapping distributions that we have described above began to happen with this analysis: unstable factor values for this factor group and for the input probability that changed radically with minor revisions to the coding scheme, failure to achieve convergence and significance, etc. The moral of the story is that the same rules of thumb that we have established for overlaps between two factors in different groups apply to overlaps between any factor and the input probability: don't push it past 95%.

5. Significance testing within factor groups. In recent variable rule work there has been a fair amount of attention paid to the significance of entire factor groups. This is no doubt due to the widespread availability of Varbrul2S and Varbrul3 which automatically calculate factor group significances through the step-up/step-down procedure. But in my experience this often leads users to neglect analytical refinements within groups.

We often start an analysis with an exhaustive list of finely discriminated environments that include every distinction we can think of that might possibly be relevant. But many or most of these distinctions prove to have no real effect on the variable under investigation. Our goal, as in any scientific endeavor, should be to achieve an analysis which is EFFICIENT, GENERAL, and POWERFUL. This means we must identify the insignificant factors which contribute nothing to the analysis and throw them out. This is the basic tenet of the scientific method known as 'Occam's Razor': always minimize the number of explanatory principles.

Practically speaking, doing this in a variable rule analysis involves a two-fold process:

1. Identifying combinations of factors that are LINGUISTICALLY more general, and
2. Checking to see if these are QUANTITATIVELY similar, in terms of their effects on the variation under investigation.

Thus if one were examining some syntactic variable for which an initial distinction between subjects, direct objects, and indirect objects had been made, it might be reasonable to try a reanalysis in which the two types of object were collapsed, provided that the probability values obtained for them were very close. A recoded data set would be created in which this factor group becomes simply subject versus object, and one would apply the standard test for significance. But what would not be reasonable would be a procedure such as the following. Say that one noticed that indirect objects and subjects had similar probability values, and then proceeded to collapse them into a single category, opposed to direct objects. In the absence of any linguistically meaningful generalization about common properties that unite subjects and indirect objects, this procedure would yield garbage, under the ancient law of computer science that states: 'garbage in, garbage out'.

Again a real-data example will serve to illuminate the issue. Table 9 presents some findings from Guy 1981 showing the effect of subject position on the variable subject-verb agreement rule in popular Brazilian Portuguese. Initially five categories were distinguished, including two separate cases where no surface subject was present. In category 2, the plural reference of the absent subject is recoverable from the complement, in sentences like *São cinco irmãos* '(They/ There) are five brothers'. In category 3 the plurality of reference is recoverable from elsewhere in the discourse, usually by continuity of reference in a particular discourse topic. (Cases where plurality of reference cannot be established by any other means are necessarily omitted from the corpus, because of an observer's paradox. See Guy 1981:243 for details.)

These two categories turn out to have very close factor values: .61 and .55. They also have a great deal in common linguistically: they constitute all and only the collectible cases with zero surface subject. Combining them into a single factor will only eliminate the secondary distinction as to how plurality of reference is established. In a reanalysis, therefore, I collapsed these two factors into one and found by means of the log-likelihood test that the distinction thus eliminated was not significant (p>.70). We can conclude that the original distinction was irrelevant to this rule, and that eliminating it yields a more general and efficient model of the variability.

Table 9: Brazilian Portuguese: Subject-verb agreement. Effect of subject position. (from Guy 1981)

	% Plural Marked	N	Probability Values		
1. Surface subject (S.S.) immediately precedes verb	74	1861	.66		.67
2. No S.S., but plurality is recoverable from complement	68	222	.61	] a	
3. No S.S., but plurality is recoverable from discourse	63	1097	.55	] a] b	.51
4. S.S. distantly precedes verb	56	597	.43	]	
5. S.S. follows verb	27	199	.27		.32

[a]difference not significant (p>.7). [b]difference not significant (p>.5)

We can also note in the original analysis that category 4 is relatively close in factor value to categories 2 and 3, at least close enough to 3 to be worth testing for significance. The linguistic generalization that can be made about all three categories is that while in categories 1 and 5 there is a surface subject ADJACENT to the verb, in 2,3, and 4 there isn't. So a further reanalysis was conducted, collapsing the original category 4 with the combined categories 2&3. Again this distinction proved insignificant (p>.50).

With these combinations made, we arrive at the most general and efficient analysis possible, shown in the last column of Table 9. A subject immediately preceding the verb favors plural-marking, a postposed subject disfavors, and a deleted or distant subject has a neutral effect.

Proceeding in this fashion we have achieved results that are intelligible and general. But suppose we had adopted the wrong approach, proceeding on grounds of quantitative similarity alone. The closest pair of probability values in the first analysis were actually factors 1 and 2, which differ in value by only .05. Collapsing them may or may not yield a statistically significant outcome, but what would the resulting combined category MEAN? What is the linguistic generalization that allows us to put together an immediately preverbal subject with one type of deleted subject, without including other deleted subjects or other preverbal subjects? To use a term from phonology, categories 1 and 2 do not form a NATURAL CLASS, regardless of their calculated numerical similarity.

This example also illustrates the point that quantitative work is not a substitute for but rather an adjunct to linguistic analysis. Varbul only performs mathematical manipulations on a set of data. It does not tell us what the numbers mean, let alone do linguistics for us. If we are asking: 'What is a good linguistic generalization?', the answer comes from our linguistic theory, not from a statistics program.

6. Independence and interaction. Varbrul analysis differs from some types of multivariate analysis in that it assumes that the various factor groups have independent effects. But users should clearly understand that THEY are not required to assume independence. On the contrary, one should be aware of the problems posed by interaction, know how to detect it when it occurs, and know what to do about it.

Discussions of the detection and treatment of interaction are to be found in Cedergren 1973 and elsewhere in the literature. To minimize repetition, I will only briefly review the points that are emphasized elsewhere, and concentrate here on some other aspects of the problem. The basic way to detect interaction with Varbrul, as Cedergren illustrates, is to look for high chi-square values in those cells where the interacting factors co-occur. All versions of the program will generate a list of chi-squares per cell in the output, and one simply examines this list for large values (ignoring ones that are artificially inflated by small cell size: n's of 1 or 2). If several such values exist, involving repeated co-occurence of a particular pair of factors, interaction is probably responsible. The chi-square is a measure of goodness of fit between the model, which assumes independence, and the data, which may have interaction. When the fit is bad, it is usually because of interaction.

Another useful technique for examining the independence of analytical dimensions is to compare the results of multiple-factor-group and single-factor-group analyses. For example, if we were looking at some phonological process that was conditioned by adjacent consonants, we could create one large factor group containing all the consonantal segments of the language, or we could have three different factor groups which classified each segment for Place, Manner, and Voice. The three-group analysis would in effect test the various distinctive features separately for their effect on the rule, while the single-group analysis would look at each segment as an unanalyzable whole, without making any claims about natural classes of segments.

A similar example involving social factors would be an analysis that had a factor group for sex and another for class (say MC vs. WC), as compared with an analysis that had a single factor group with four factors, one for each subgroup of the population: {WC males, MC males, WC females, MC females}.

The independence 'test' in this procedure involves comparing the probability values obtained by the two analyses. If the single-group figures map out in a table systematically along the lines predicted by the multiple-group analysis, there's no interaction. But if some segment or social subgroup does not line up as expected based on the 'feature' (multiple-group) analysis, then there's a possibility of interaction. This would arise in our class-by-sex example if the following case occurred. Suppose that the two-group analysis showed higher probabilities for males than females, and for working class speakers than middle class. We would then expect that in the single-group analysis, the factor value for working class males should be greater than that for MC males, among other things. If this were not true, if say, these two figures were equal or reversed, then it would be likely that class and sex were not independent.

An example will help to clarify these issues. In Table 10 are listed the probabilities obtained in Guy 1981 for the effect of a following consonant on the deletion of final -s. The feature analysis shows that voiced consonants favor deletion more than voiceless, and that there is a place effect involving a peak of deletion for dentals and a minimum value for velars.

If the results of the segmental analysis perfectly reflected the results of the feature analysis, then each voiced segment should have a higher value than its voiceless counterpart at the same place of articulation. Also the values in each row in the segmental analysis table should be ordered dental>labiodental>labial>velar. These relationships hold for all the values in the table except one, the segment /g/. It is LOWER than its voiceless counterpart /k/, when it should be higher. Based on the overall pattern, we would expect /g/ to have a

value between .35 and .40, rather than .12. This means that there may in effect be some interaction between the features [+voice] and [velar]. The statistical consequence is that the feature analysis proves to be significantly WORSE on the log likelihood measure than the segmental analysis ($p<.05$). (The feature analysis is the one that is worsened by interaction, because it is more highly constrained. It tries to assemble natural classes of segments, whereas the segmental analysis allows the factor value of each segment to float freely, unconstrained by what other phonetically similar segments do.)

Table10. Brazilian Portuguese: Final -s deletion. Effect of following consonant. (Probabilities of deletion, from Guy 1981.)

Segmental Analysis

b	v	d	g
.47	.63	.72	.12
p	f	t	k
.38	.45	.53	.27

Feature Analysis

Voicing:	
voiced segments:	.58
voiceless segments:	.42

Place:			
labial	labiodent.	dental	velar
.50	.54	.64	.32

So having identified a potential interaction here, what should we do about it? There are two answers to this question, depending on one's purpose. Analytically, there is an accepted device for treating interaction within the Varbrul framework, by means of an interaction term or 'exception feature.' This is accomplished by adding an extra factor group to the feature analysis, say {x,y}, where 'x' is coded in every cell containing [+voice] AND [velar] (i.e. every cell containing the exceptional case of /g/), and 'y' is coded in all other cells. The resulting analysis will assign the exceptionally low value of /g/ to the 'x' factor.

The reason one does this is that the values obtained WITHOUT using an exception feature are going to be distorted by the interaction. If /g/ is exceptionally low, then the values calculated for [+voice] and [velar] will be too low also, in order to achieve a better statistical fit between model and data. When the exceptional nature of /g/ is accomodated by the inclusion of an interaction term in the model, a more accurate estimate of the values of the other factors can be obtained.

The second answer to the question 'what to do' is a theoretical one—that is to try and prove that the results are wrong or suspect, and that there is really no interaction. One does this when one is confident on theoretical or other independent grounds that there should not be any interaction between the terms in question, and is therefore suspicious of the finding. In the present case I could find no reasonable explanation whatsoever for why /g/ should behave peculiarly in conditioning -s deletion, so I was highly suspicious of this result. If it were a case of interaction between some social parameters, like, say, sex and occupation, I would not be the least bit suspicious. But this result was theoretically unlikely, so I sought alternative explanations.

A possible answer was found in the figures for data quantity in this case. The N's for Table 10 are found in Table 11. There we see that the offending segment /g/ has by far the smallest number of tokens supporting the analytical results. So in this case, despite a statistical finding of significance at the .05 level, I decided not to worry about it. One should remember that at this level of significance one still expects one case in twenty to be produced by purely random fluctuations in the data. In other words, there is one chance in twenty that the anomalously low value for /g/ arose because it happened quite randomly that the set of final -s's followed by words beginning with /g/ that we encountered included just

by chance a smallish number of deletions. In this instance my theoretical expectations were strong enough that I interpret this to be that one case in twenty. This view can of course be tested by replication: collect more cases, say another 75 or 100 tokens of final -s before /g/, and see if the same low rate of deletion prevails. If so, the phenomenon is real and needs to be accounted for; if not, my suspicions are confirmed. I leave this task to an interested reader.

Table 11. Brazilian Portuguese: Final -s deletion.
Numbers of tokens, by following consonant. (From Guy 1981)

b	v	d	g
122	278	736	75
p	f	t	k
497	314	498	961

7. Conclusions. Varbrul is a powerful and extremely useful tool for the analysis of linguistic variation. As with any tool, its utility is enhanced by an understanding of its operations and its limitations. But one should never lose sight of the fact that, in the final analysis, it is only a device (albeit a sophisticated one) for manipulating the data. It does not discern patterns, make generalizations, or explain findings. All that is up to you.

References

CEDERGREN, HENRIETTA J. 1973. The interplay of social and linguistic factors in Panama. Dissertation, Cornell University.

_____, and DAVID. SANKOFF. 1974. Variable rules: Performance as a statistical reflection of competence. Language 50.233-55.

GUY, GREGORY R. 1975. Use and applications of the Cedergren/Sankoff variable rule program. Analyzing variation in language, ed. by Ralph W. Fasold and Roger W. Shuy, 59-69. Washington: Georgetown University Press.

_____, 1981. Linguistic variation in Brazilian Portuguese: Aspects of the phonology, syntax, and language history. Dissertation, University of Pennsylvania.

LABOV, WILLIAM. 1966. The social stratification of English in New York City. Arlington, VA: Center for Applied Linguistics.

_____, 1969. Contraction, deletion, and inherent variability of the English copula. Language 45.715-762.

POPLACK, SHANA. 1980. The notion of plural in Puerto Rican Spanish: Competing constraints on /s/ deletion. Locating language in time and space, ed. by William Labov, 55-68. New York: Academic Press.

ROUSSEAU, PASCALE. 1983. A versatile program for the analysis of sociolinguistic data. Université de Montréal, Centre de Recherche de Mathematiques Apliquées, CRMA-1163.

_____, and DAVID SANKOFF. 1978. Advances in variable rule methodology. Linguistic variation: models and methods, ed. by David Sankoff, 57-69. New York: Academic Press.

SANKOFF, DAVID. 1975. VARBRUL version 2. Unpublished manuscript.

_____, and WILLIAM LABOV. 1979. On the uses of variable rules. Language in Society 8.189-222.

WEINREICH, URIEL, WILLIAM LABOV, and MARVIN HERZOG. 1968. Empirical foundations for a theory of language change. Directions for historical linguistics, ed. by Winfred Lehmann and Yakov Malkiel, 97-195. Austin: University of Texas Press.

Syntactic, Semantic, and Pragmatic Influences on Judgments of Grammaticality

Alan and Allison Hudson-Edwards
University of New Mexico and Albuquerque Public Schools

Introduction. In a well-known paper delivered to the Chicago Linguistic Society in 1969, Elliott, Legum, and Thompson (ELT) argued for the significance of systematic variation in judgments of grammaticality as data to which any complete theory of language must be accountable. Although modest in its dimensions, the paper justifiably gained rapid and widespread recognition because of the attention it drew to the study of syntactic variation (in contrast with the earlier emphasis on phonological variation), because of the use of implicational scaling as a technique for capturing patterns of interspeaker variability, and because of the questions which it inevitably raised regarding the proper object of grammatical description.

Using four variants of the basic sentence *Sophia Loren was seen by the people while enjoying herself*, ELT found that sentences were judged more grammatical when the reflexive pronoun in the subordinate clause was coreferential with the surface structure subject of the main clause. Compare sentences 1 and 2 in Table 1 with sentences 3 and 4 respectively.

Table 1. Grammaticality Ratings from Elliott, Legum & Thompson 1969
(calculated from published data, lower = more grammatical)

1. Sophia Loren was seen by the people while enjoying herself.	1.6
2. The people saw Sophia Loren while enjoying themselves.	1.5
3. Judy was seen by the people while enjoying themselves.	2.6
4. The people saw Karen while enjoying herself.	3.1

The results for active vs. passive voice are more equivocal. Although ELT did not test for statistical significance, there probably is no replicable difference in grammaticality between active and passive voice when the reflexive pronoun is coreferential with the subject of the main clause (Compare sentences 1 and 2), although there is a difference, probably significant, in favor of the passive voice when the reflexive is coreferential with a main clause non-subject (Compare sentences 3 and 4).

Intuitively, ELT's findings make good syntactic sense, particularly with regard to the preference for sentences in which the reflexive pronoun is coreferential with the surface structure subject of the main clause. However, given the somewhat strange selection of stimulus sentences, we were prompted to ask whether the differences in judgments of grammaticality might not have been due to considerations of a semantic, pragmatic, or thematic nature, instead of, or in addition to, the kinds of syntactic constraints with which ELT were originally concerned. The present study, therefore, was designed to test for the relative roles of syntactic, semantic, and pragmatic influences on perceptions of grammaticality.

Method. Sixteen stimulus sentences were administered on paper to a total of 303 students in basic linguistics courses at the University of New Mexico. The respondents, all native speakers of English, were asked to rate the grammaticality of each sentence on a scale from one to four, the lower end of the scale representing complete grammaticality and the upper end ungrammaticality. Unlike ELT, the subject and non-subject of the main clause in each sentence were both anonymous, thus eliminating the celebrity of either one as a possible confounding variable. For similar reasons, all subjects and non-subjects were