Statistics in Sociolinguistics

D. Sankoff

1. Introduction

Many areas connected with sociolinguistics in which quantitative data play a role, have seen the application of statistical methods, both traditional (experimental design, sampling, estimation, hypothesis testing) and heuristic (clustering, scaling). Some of these are discussed in other entries (see *Social Psychology*; *Sociometry*; *Attitude Surveys: Question and Answer Process*; *Scaling*; *Multidimensional Scaling*). It is within variation theory, however, where social considerations are most intimately connected to grammatical questions, that a specifically sociolinguistic protocol for statistical analysis has been developed and widely adopted. This article will survey the justifications for this approach and sketch how it has been applied to various types of linguistic problems.

2. Epistemological Concerns

Linguistics is unique among scientific disciplines in that its practitioners generally do not require statistical methodology and are not constrained by statistical criteria of validity. Linguists traditionally agreed that the grammatical structure of a language consisted, in large measure, of discrete entities or categories whose relationships and co-occurrence constraints were qualitative in nature and shared by all speakers in the speech community. These structures could then be deduced by analyzing and comparing test utterances elicited from, or intuited by, any native speaker of the language (e.g., linguists serving as their own data source), without need for any statistical apparatus.

It is only since the groundbreaking work of William Labov in the late 1960s that any concerted attempt has been made to investigate questions of central interest to linguistic theory using statistics (Labov 1969). He made credible to many linguists the idea that two or more different articulations of a given phonological form may occur in the same word or affix, in the same contexts, without affecting the denotational value (referential meaning) of a lexical item, or the syntactic function of an affix or particle. Which form will occur at a given point in time can only be predicted in terms of a probabilistic model, whereby the effects of the linguistic and extralinguistic context can be ascertained with accuracy, but where the output of the analysis remains just a probability. The choice of form always contains a component of pure chance, though this is precisely delimited.

Alternation among forms also occurs at the syntactic and lexical levels, though the scope of syntactic equivalence and lexical synonymy are the subject of much debate. The distributionalist tradition of linguistics is based on the uniqueness of the form-function relationship, something variationism rejects as an unwarranted assumption. Analysts may well identify, *upon reflection*, differences in connotation among synonyms or among competing syntactic constructions, whether in isolation or in context, but there is no reason to expect these differences to be pertinent every time one of the variant forms is used. Indeed, underlying the study of syntactic variation within a framework similar to that of phonological variation is the hypothesis that for certain identifiable sets of alternations, these distinctions come into play neither in the intentions of the speaker nor in the interpretation of the interlocutor. We say that 'distinctions in referential value or grammatical function among different surface forms can be neutralized in discourse' (Sankoff 1988). This is the source of the phenomenon of 'equivalence' and the justification of the syntactic variable that have so preoccupied sociolinguists. It is the fundamental mechanism of non-phonological variation and change.

The formal models of grammatical theory have discrete structures of an algebraic, algorithmic, and/or logical nature. Such structures often involve sets of two or more alternate components, such as synonyms, paraphrases or allophones, which the analyst may determine to be carrying out identical or similar linguistic functions. By allowing a degree of randomness into the choice between such alternates, the grammatical formalisms are converted into probabilistic models of linguistic performance susceptible to statistical study (Cedergren and Sankoff 1974).

3. The Sample

Language modeled as generated by probabilistic grammars is most appropriately studied by the data of natural discourse: sustained fluent sequences of connected utterances. The construction of a sample of natural speech is very different from sampling for sociological questionnaire administration, for psychological experimentation, or for educational testing, since one generally cannot predict when or how often the linguistic phenomenon under study will occur in the flow of conversation. Hence the sample usually involves relatively few speakers (20 to 120), carefully chosen to represent the diversity of linguistic behavior within the community being studied, with a large volume of material tape-recorded from each speaker. This material is

transcribed in computer-accessible form, and is systematically scanned for occurrences of the words, sounds, or grammatical structures of interest. Usually, the same data set (the 'corpus') can be used for many different studies, since it is representative of all the structures and usage of natural speech (Sankoff and Sankoff 1973).

4. The Research Question

Corpus research (see *Corpus Linguistics and Sociolinguistics*) may provide examples or counter-examples as an adjunct to theoretical arguments. But what is equally important, and often more so, are quantitative patterns of occurrence relatively inaccessible to introspection or even testing methodology. Regular, complex relationships may exist at the quantitative level among a number of structures, but upon introspection, all we can say is that they are all simply 'grammatical.' Quantitative regularities may be vaguely guessed at through introspection, but may not be characterized with anything like the precision with which intuition-based methods can establish categorical relationships.

The quantitative facts are not minor details of linguistic behavior. Universal hierarchies and co-occurrence constraints not manifested in terms of grammaticality versus ungrammaticality for a given language are nonetheless often present in clear and well-developed form in usage frequencies. The classical examples are constraint hierarchies for the expression of certain allophones (or the application of optional phonological and morphophonological rules), but it is also true of syntax, in the study of variable rule order, optional movement or deletion rules, and in preferences among semantically or functionally equivalent phrase structures.

Moreover, it is these variable aspects of grammatical structure which are always the locus of linguistic change (see *Sociolinguistics and Language Change*). Change virtually always requires a transitional period, often very lengthy, of variability, competition among structures and divergence within the speech community. The detailed nature of linguistic change and of its synchronic reflex—dialect differentiation—cannot be understood without coming to grips with quantitative relationships.

The tools for studying these relationships are necessarily very different from those used in theoretical linguistics. Frequency counts of forms in contexts are not just quantitative refinements of judgments of grammaticality and have even less to do with acceptability. Counts of 0 percent are analogous to judgments of ungrammaticality, but not identical to them. Non-occurrence does not necessarily indicate a prohibited form. It may simply be the result of a complex combination of features, which could be perfectly grammatical but unlikely to appear in any reasonably sized corpus. Conversely, intuitively ungrammatical forms may appear systematically and at a non-negligible rate in spontaneous speech through the interaction of the grammatical facility with processing constraints.

5. The Linguistic Variable

The key concept underlying variationist sociolinguistics is the 'linguistic variable.' A well-studied example is the pronunciation or deletion of syllable-final /s/ that characterizes most varieties of Caribbean Spanish. Another example, from spoken English, is the copular verb *be*, which occurs as the contracted variant (*John's a doctor*, *we're coming*, *I'm at home*) or the full variant (*John is ... we are ... I am ...*). A third example involves *th*, as in *this* and *think*, which usually have an interdental fricative pronunciation, but which are also pronounced at least occasionally by speakers of most varieties of English as stops (*dis*, *tink*). A fourth example is the alternation of future and periphrastic future tenses (*You will hear about it* versus *You are going to hear about it*).

6. The Model

The choice of one variant or another of a binary linguistic variable can be heavily influenced by a wide range of factors, including the phonological and syntactic context in which it occurs, the topic of conversation, the degree of situational or contextual formality, idiosyncratic tendencies of the speaker, and the identity of the hearer(s). These factors, however, usually cannot account for all the variability in the data, and so a probabilistic model is set up to evaluate their influence:

$$\log p/(1-p) = a_i + a_j + \cdots + a_k \qquad (1)$$

where p is the probability that a particular one of the two variants will be chosen and $1-p$ the probability of the other, in the context containing linguistic or extralinguistic features labeled $i, j, \ldots, k$. In this 'logistic-linear' model $a_{i\ \text{represents}}$ the effects on the choice of variant of feature i in the context of the variable. If p represents the probability of contraction in the example of copula given above, and if feature 1 and feature 2 indicate that the sound preceding the copula is a vowel or consonant, respectively, while feature 3 and feature 4 indicate that the grammatical category following the copula is an adjective or a noun phrase, and feature 5 and feature 6 indicate informal and formal speaking styles, then a_1, a_3, and a_5 will be high (each ca. 1.0) while a_2, a_4, and a_6 will be low (each ca. -1.0). Thus the *is* in *John is the leader*, as uttered in a formal context is far less susceptible to contraction ($p = 0.05$) than the *are* in *You are bad*, spoken informally ($p = 0.95$).

With sociolinguistic data, it is conventional to include one feature, called the mean or corrected mean, in formula (1) for all contexts. The other

features fall into disjoint 'factor groups.' Usually one or another feature from each factor group appears in every context, and no two features from the same factor group can co-occur. The mathematical structure of (1) requires us to impose some constraint on the feature effects within each factor group, such as that the sum of the effects be zero. Without this the mean cannot be estimated uniquely.

7. The Data

For statistical analysis, the speech sample is scanned for occurrences of one or the other variant of a variable, and each occurrence is recorded along with the features (or factors) present in the context. Each observed combination of contextual features is considered to define a data cell for the analysis, where the number r of occurrences of one of the variants, compared with the total occurrences n in the cell, is assumed to be a binomial random variable with parameters n and p.

Because the distribution of the data among the cells cannot be controlled when a natural speech sample is used, and because many different factors may influence the probability p, the final 'design' is often a high-dimensional array with many or even most of the possible cells empty ($n = 0$), and the data is distributed very unevenly among the others. Estimation methods based on sum-of-squares approximations such as multiple regression and analysis of variance are thus inappropriate and the parameters must be estimated using exact maximum likelihood methods. Many statistical computing packages have the capability of carrying out this type of analysis, although most of the linguistic work has made use of one or other version of the 'variable rule' program (Rand and Sankoff 1990, Robinson et al. 2001). Elimination of statistically irrelevant influences can be assured by a multiple regression type of analysis with a stepwise selection of significant factors. For example, in expressing future time in the French spoken in Montreal, if a verb is negated, this has a statistically very significant effect in reducing the use of the periphrastic future. Elevated socioeconomic status of the speaker is also a significant factor, while neither the nature of the subject of the verb nor the age of the speaker has a significant effect.

8. Detecting Heterogeneity

A major preoccupation in sociolinguistics is the degree of homogeneity of the speech community. Do all speakers in the community share a common model of type (1)—possibly involving a single parameter to account for individual differences—or might different individuals or different segments of the community each have a substantively different model of type (1)? A way of answering this question is based on a dynamic clustering procedure. An initial (random) partition into k groups of the speakers in the speech sample is made, followed by an estimation of the parameters in k versions of model (1), a separate model for each group. Speakers are then reassigned to groups according to which model they 'fit' best, using the likelihood criterion. Further iterations are carried out on the estimation and reassignment procedures until they converge. The significance of the analyses for each of $k = 2, 3 \cdots$ can be tested based on the increase in likelihood with the increase in k, compared with the number of additional parameters estimated.

Thus, in expressing an indefinite referent, Montreal French speakers fall into two groups according to how they vary between *on* 'one' and *tu* 'you.' In one group, speakers have a high rate of *on* usage in conveying proverb-like sentiments and in a certain class of syntactic construction while the other group shares the former but not the latter (syntactic) effect (Rousseau and Sankoff 1978a).

9. Implicational Scales

Data on a linguistic variable are sometimes given as a two-dimensional array, where each row represents a different speaker or speech variety and each column represents a different linguistic or sociolinguistic context. In the simplest version, each cell of the table contains a 0 or a 1, indicating which of two variants of a linguistic variable is expressed by the given speaker in the given context. The problem is to reorder the rows and columns so that every row consists of a series of 0's to the left followed by a series of 1's to the right (i.e., no 0 is to the right of a 1 in a row) and every column consists of a series of 0's at the top followed by a series of 1's on the bottom (i.e., no 0 is below a 1 in a column), as in Table 1.

If an implicational scale can be established, it represents a highly economical representation of the how the set of speakers and the set of contexts are organized in terms of this linguistic variable.

It is not always possible to arrange data into an implicational scale, so various somewhat arbitrary measures of scaling have been proposed, to assess to what extent a data set is 'scalable,' based on the minimum possible number of 'scaling errors.'

In the general case, instead of 0's and 1's, each cell contains a fraction r/n, representing the successes divided by the total trials of a binomial experiment (or the uses of one of the variants divided by the total occurrences of the variable). The problem is then to find row and column permutations (or relabelings) such that in the relabeled matrix, r/n is nondecreasing along both rows and columns, as in Table 2. This type of analysis is of interest when in most cells $r/n = 0$ or $r/n = 1$, so that all except a few 'transitional' cells correspond to the 0's and 1's of the basic analysis.

Table 1. Simple implicational scale depicting variant use (0 versus 1) in 5 contexts by 6 speakers.

	context 1	context 2	context 3	context 4	context 5
speaker 1	0	0	0	0	1
speaker 2	0	0	0	0	1
speaker 3	0	0	0	1	1
speaker 4	0	0	0	1	1
speaker 5	0	1	1	1	1
speaker 6	1	1	1	1	1

Table 2. General implicational scale depicting one variant's frequency r divided by total occurrences n of the variable.

	context 1	context 2	context 3	context 4	context 5
speaker 1	0/3=0	0/1=0	0/5=0	1/6	2/4
speaker 2	0/10=0	0/3=0	0/3=0	5/12	8/9
speaker 3	0/2=0	0/1=0	0/1=0	6/6=1	10/10=1
speaker 4	0/7=0	0/4=0	0/2=0	2/2=1	3/3=1
speaker 5	0/6=0	3/4	8/10	6/6=1	6/6=1
speaker 6	2/7	7/8	8/9	1/1=1	9/9=1

It might seem that a logistic-linear model such as (2) would not seem capable of accounting for data which form an implicational scale since p could not be 0 or 1 for any cell as long as the speaker effect a and the row effect b are finite.

$$\log p/(1-p) = a + b \qquad (2)$$

In other words, model (2) could only give an approximate account for nonvariable behavior (i.e., when $r/n = 1$ or $r/n = 0$). In properly defined maximum likelihood estimation for data such as those in Table 2, however, the estimates for a and b in (2) remain undefined, but the estimates of p in each cell are well defined, and can take on values 0 and 1 where the data predicts nonvariable behavior (Rousseau and Sankoff 1978b). This forms the basis for an integrated logistic-linear/implicational scale analysis, including a principled basis for rejecting data which cause 'scaling errors,' since these data turn out to be outliers in terms of their extremely low likelihood under a maximum likelihood analysis of the data set (Sankoff and Rousseau 1979).

10. Rare Variants

Variation theory sometimes encounters a situation, especially in syntax, where one of the two variants is extremely rare, and may not even occur once in the entire corpus at hand. This variant may well represent an important phenomenon, however, such as an incipient change or a necessary condition of some other rare contextual feature. Examples of the rare variant may only be collected through what is effectively participant observation, writing down sentences as they are heard in daily conversation, on the radio or television, or read in the written media.

We then effectively have two corpora, the conventional one containing no examples of the rare variant but sufficient examples of the usual variant distributed among a sample of contexts, and the new corpus containing exclusively examples of the rare variant, presumably in a representative sample of the contexts in which it occurs. Though it may seem somewhat counterintuitive, variable rule analysis based on formula (1) can be carried out as normal on the combined data set, as if all the occurrences originated in the same corpus (cf. Bishop 1969). The only difference is that the estimate of mean or corrected mean has no meaningful interpretation, since this would depend strongly on the (unknown) total amount of conversation from which the rare variant examples were extracted. The feature effect estimates, however, have their normal interpretation and are not affected by the dual origin of the data.

11. Multiple Variants and Rule Order

In the preceding sections, we have only discussed choice variables, which are dichotomous, although the factor groups could contain any number of factors. There are many cases, however, where a linguistic choice may be perceived as involving three or more alternatives simultaneously. Generalization of the mathematical treatment to this context offers no conceptual difficulty, only a number of complications in the formulae.

For example, consider the case of three variants A, B, and C, with probabilities p, q, and r in a given

context. The logistic-linear model becomes:

$$\begin{aligned} \log(p/q) &= a_i + a_j + \ldots + a_k \\ \log(q/r) &= b_i + b_j + \ldots + b_k \\ \log(r/p) &= c_i + c_j + \ldots + c_k \end{aligned} \tag{3}$$

Actually, the third equation is equivalent to the sum of the other two, and hence is redundant. In general, if there are n variants, then $n-1$ equations suffice to define the model.

When there are multiple variants, it is not always clear whether they should be considered as having been generated simultaneously, or whether certain distinctions among the variants are decided before others. This part of the classical problem of rule ordering can be studied through variable rule analysis. The other part of the problem, the question of what forms underlie what others, cannot be approached through statistical analysis of conditioned choice data.

For example, consider a hypothetical variable with four surface forms (variants 1 through 4). Mathematically speaking, there are 184 ways of generating these by different rule order schemes from a common underlying form, though many of these may be implausible. Two schemes are depicted in Fig. 1.

The first rule in scheme (a) can be analyzed by a two-variant variable rule, since we know how many times in each context the choice was made to rewrite (the number of surface occurrences of variants 2, 3, and 4) and how many times it was not (the number of occurrences of variant 1). Similarly, the choices represented by the second rule can be analyzed using the number of variant 2 versus the number of 3 and 4. And the third rule is based simply on the number of 3 versus the number of 4. In scheme (b), the two-pronged rule represents a simultaneous three-way choice between variants 1, 2, and 3, and can be analyzed using the number of variant 1 versus (2 and 4) versus 3.

The likelihood of each of the 184 schemes can be calculated as the product of the maximum likelihoods of the individual choice processes represented by its rules. We can then invoke the principle of maximum likelihood to infer the best scheme. Suppose that analysis of data showed that (b) in Fig. 1 is the most likely scheme. Then there would be at least five other

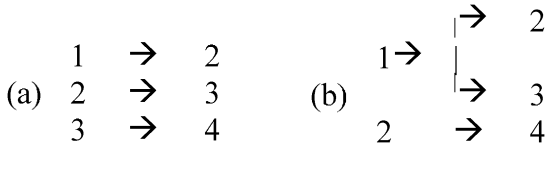

Figure 1. Rule order schemes: (a) some occurrences of underlying variant 1 rewritten as 2, then some of these 2 rewritten as 3, then some 3 rewritten as 4; (b) some 1 rewritten as either 2 or 3, then some 2 rewritten as 4.

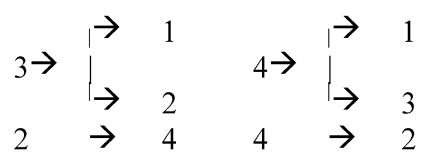

Figure 2. Two schemes with the same likelihood as scheme (b) in Fig. 1.

schemes, such as the two in Fig. 2, which must have exactly the same likelihood, since they imply exactly the same choice processes. They differ only according to which forms underlie with others.

The implication of this equivalence for the problem of rule ordering is that this problem can be decomposed into two aspects. One is the identity of the underlying form or, more generally, which variants give rise to which other, and the second aspect is the most likely arrangement of the variants into a treelike, or hierarchical classification, of which there are only 26 different possibilities compared to the 184 schemata in the example. The data on variant occurrences in context do not bear at all on the first aspect, but do allow us to use statistical means, namely the comparison of the likelihoods of the different schemata, to infer the hierarchy. Thus only the order in which the distinctions among subsets of variants are drawn can be studied statistically, and not the question of underlying forms.

This methodology has been used most extensively in analyzing the variation in the pronunciation of the word-final consonants /s/, /n/, and /r/ that characterize most varieties of Caribbean Spanish (Sankoff and Rousseau 1989).

12. Discussion

Much of this article has been phrased in terms of alternations between forms that are related in some theoretical account of phonology or syntax. However, where statistical regularities are found in linguistic performance, they are important as properties of language independent of whether they are consequences of:

(a) the physiology of articulation, in phonology;
(b) processing considerations in syntax;
(c) social or biological universals, as in the competition of tense and aspect inflections with periphrastic constructions based on verbs for standing, sitting, going, etc., or in the competition of modals with verbs for volition, ability, desire, etc.;
(d) panlinguistic typological tendencies ('parameters') which may or may not be coded in some innate form on the individual level; or
(e) some punctual actualization of the individual's grammatical facility.

There are many types of causes of statistical regularity, and which one or ones are pertinent to a given linguistic pattern remains an empirical question.

See also: Sociolinguistic Variation; Sociometry; Multidimensional Scaling.

Bibliography

Bishop Y M 1969 Full contingency tables, logits and split contigency tables. *Biometrics* **25**: 119–25

Cedergren H J, Sankoff D 1974 Variable rules: performance as a statistical reflection of competence. *Language* **50**: 333–55

Labov W l969 Contraction, deletion, and inherent variability of the English copula. *Language* **45**: 715–62

Rand D, Sankoff D 1990 *GoldVarb: A Variable Rule Application for Macintosh*. Centre de recherches mathématiques, Université de Montréal

Robinson J, Lawrence H, Tagliamonte S 2001 *GOLDVARB 2001 for Windows*. Department of Language and Linguistic Science, University of York

Rousseau P, Sankoff D 1978a A solution to the problem of grouping speakers. In: Sankoff D (ed.) *Linguistic Variation: Models and Methods*. Academic Press, New York

Rousseau P, Sankoff D 1978b Singularities in the analysis of binomial data. *Biometrika* **65**: 603–8

Sankoff D 1988 Sociolinguistics and syntactic variation. In: Newmeyer F (ed.) *Linguistics: The Cambridge Survey*. *IV: The Socio-Cultural Context*. Cambridge University Press, Cambridge

Sankoff D, Rousseau P 1979 Categorical contexts and variable rules. In: Jacobson S (ed.) *Papers from the Scandinavian Symposium on Syntactic Variation*. Almqvist & Wiksell, Stockholm

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

Sankoff D, Sankoff G 1973 Sample survey methods and computer-assisted analysis in the study of grammatical variation. In: Darnell R (ed.) *Canadian Languages in their Social Context*. Linguistic Research Incorporated, Edmonton

Please check the Quality of Figs 1 and 2