Chapter 4

Probabilistic Sociolinguistics: Beyond Variable Rules

Norma Mendoza-Denton, Jennifer Hay, and Stefanie Jannedy

4.1 Overview

In this chapter we outline issues facing quantitative approaches in contemporary variationist sociolinguistic theory, surveying trends that led scholars (1) to reject intuitive and categorical descriptions of language data and (2) to use frequency-based and probabilistic approaches to the modeling of language variation. We discuss the importance of linguistic and social contexts in the description of variable linguistic behavior by analyzing our data on the monophthongization of the diphthong /ay/ in the speech of the popular African-American talk show host Oprah Winfrey. We compare VARBRUL (variable rule-based logit analysis) results with CART (classification and regression trees) results to highlight the strengths and weaknesses of different tools in the modeling of probabilistic phenomena. Implications for the theory of sociolinguistic variation and for models of cognition are emphasized throughout. We advocate a usage-based account for both linguistic processes and social identity construction. Such an account allows for the continuous and incremental updating of mental representations on the basis of new input, and it synergistically captures advances in probabilistic linguistics and in social identity construction theory. Social identities are transmitted simultaneously with linguistic structures, and as such they represent dynamic processes, continuously negotiated in interaction.

4.2 Background and History

4.2.1 Introduction

The study of sociolinguistic variation has faced different problems from other fields of linguistics. While other branches of quantitative linguistics

have competed with schools of intuitive and categorical thinking (Bod, this volume), sociolinguists have always started from empirical premises. The very first statistically sophisticated studies that were conducted in a modern sociolinguistic framework laid the foundation for debates on statistical modeling within this field. Past debates within sociolinguistics have included the search for a unified statistical model and tools (Bickerton 1971; Sankoff and Rousseau 1974); the interpretation of correlational statistics linking social structure to linguistic forms, especially in the field of language and gender (Eckert 1989; Labov 1990; Cameron 1990); and the positing of alternative models for the diffusion of change through a population, such as the implicational scale versus quantitative model debate (Bickerton 1973; Romaine 1985; Rousseau 1989; see summary in Rickford 2001). Several of these debates have accorded privileged status to questions of how to model the mathematics of sociolinguistics, while paying short shrift to cognitive issues of the mental representation of linguistic categories and of social processes. Recent work by Mendoza-Denton (2001) and Eckert (1999) has pointed out that advances within social theory and the evolution of understanding of sociolinguistic processes challenge researchers to move beyond viewing social categories as static, relegating them to simple decisions made by the analyst prior to data analysis. Primary questions now surfacing are: How do social categories emerge from the distribution of data? How do abstractions such as ethnicity and gender emerge from the many different ways that speakers have of fashioning themselves as classed, gendered, or ethnic social agents? Although some of the current methods (such as VARBRUL and CART) constrain researchers in selecting discrete variables within sociodemographic categories (coding tokens for age, ethnicity), we propose utilizing a variety of techniques (including discourse and conversation analysis) to more closely examine specific instances of variables and the contexts of their use to determine how social meaning is constructed.

Exemplar theory, a frequency-based model emerging in areas such as phonology and morphology (Pierrehumbert, this volume), can lead the way to unification with social-theoretic understandings of the role of innovative social actors in communities of practice. In exemplar theory, categories are not preexisting, but are established as dynamic (continuously and incrementally updated) generalizations of linguistic data over increasingly abstract domains. The robustness of the categories depends on frequency of the input that can be classified under that category, and on the recency of the stimulus.

There is a groundswell of evidence that much social information is carried in moment-to-moment performances by key individuals—icons—in local communities (Eckert 1999; Labov 2001; Mendoza-Denton 2001; Schilling-Estes 2001). Performances by these social brokers in the linguistic marketplace are subject to the same cognitive constraints of robustness and frequency that underlie other areas of symbolic manipulation.

After reviewing some of the early sociolinguistic literature on variation and on the variable rules framework, we present an extended example analyzing a socially iconic speaker—Oprah Winfrey—with two statistical modeling techniques, supplemented with discourse analysis, showing how her use of specific variants contributes to the construction of her linguistic style.

4.2.2 Against Intuition

Sociolinguistics explores the social correlations of patterns of human linguistic behavior at all levels of grammar, ranging from phonology and syntax to semantics and discourse. The quantification of performance data to explore and explain speakers' linguistic competence in social situations has been a staple of the sociolinguistic paradigm. Unlike the methods used in some other areas of linguistics, those deployed by sociolinguists are empirical in nature and require the modeling of quantitative patterns to draw conclusions about speaker competence. It is not assumed that linguistic innovation, nuances in speech patterns, or variants of lexical choice are in free variation. Rather, they are manifestations of the subtle patterning and interaction of linguistic and social competence.

A speaker has choices to make when selecting which words to use in crafting a sentence, whether to release a word-final stop, or whether to raise a high vowel to display more extreme formant values. These choices carry social meaning at the moment of utterance, and the gradual cumulative steps of innovators may lead to category shifts with the power to rearrange entire linguistic systems. Through the analysis of historical records we gain insight into the succession of linguistic changes, such as those precipitated by the English Great Vowel Shift. Historical evidence and contemporary recordings can be used to show the gradualness of these changes, the lexical diffusion of their carrier items through the population, and their continuing consequences in current structural reorganizations, such as the Northern Cities Chain Shift in the United States (Eckert 1989; Labov 2001).

Sociolinguistics is concerned with capturing not only patterns of change, but also variation across speakers of different speech communities, among speakers in a single speech community, and in the speech of individuals. Variability follows the twin constraints of (1) being conditioned by language-internal factors and (2) participating in processes of social semiosis—a dual meaning-making system par excellence. Because there is little room in generative linguistic frameworks to explore and explain either noncategorical changes or stable variation, much work in that vein has been devoted to describing the endpoints of changes, variability being dismissed as randomness or noise. Categorical descriptions of language data ignore the triggers and mechanisms of variability, their social motivation, and the productivity of such linguistic patterns.

As far back as 1937, Bronislaw Malinowski outlined a view of the essential dilemma facing linguistics:

> ... whether the science of language will become primarily an empirical study, carried out on living human beings within the context of their practical activities, or whether it will remain largely confined to deductive arguments ... (1937, 63)

This chapter will argue that current quantitative models of language behavior may still benefit from further investigation precisely of the form that Malinowski advocated: carried out on living individuals in the course of practical activity, shedding light on both linguistic form and questions of social structure.

Hymes exhorted his linguistic contemporaries to take up research in a nascent field called sociolinguistics, the goal of which was to "identify rules, patterns, purposes, and consequences of language use, and to account for their interrelations" (1974, 71). The definitional core of this field was and remains a theoretical concern for the interrelationship and the codependence between components of linguistic structure and of social structure. Why is this inherently a probabilistic problem? Sociolinguists commonly understand the *linguistic variable* as "a construct that unites a class of fluctuating variants within a language set" (Wolfram 1991, 23), reflecting a decision point at which a speaker chooses between alternative ways of saying the same thing.

The central probabilistic sociolinguistic questions then become: What factors affect a speaker's decision to use one variant over another? How can we best model the simultaneous influence of these linguistic and social factors at that particular decision point? How does the use of a particular linguistic variant reflect social membership? And what can the distribu-

tion of alternative forms in the social landscape reveal about the internal synchronic and diachronic workings of linguistic structure?

We take the multiple determination of variables as a given: it is not only the internal organization of linguistic structure (i.e., phonological context) that shapes variation, but also social settings and characteristics of speakers, all operating in concert and reflected in language (cf. Bayley's (2001) principle of multiple causation).

Labov (1966, 1972) showed that through the frequencies of the various phonetic manifestations of underlying phonological /r/, New Yorkers displayed finely tuned linguistic performance reflecting social classes, ethnic groups, and even such subjective factors as the level of formality in the speech situation. Rhoticity, the presence or absence of a pronounced syllable-coda /r/, varied in predictable and replicably measurable ways. However disparate their informal production, New Yorkers demonstrated their orientation to the current rhotic standard by exhibiting variation wherein formal speech was always more rhotic than informal speech, across all social classes and across all "styles" (word list, reading passage, formal interview, unstructured interview). Figure 4.1 illustrates class stratification in New York City as reflected in a linguistic variable. The vertical axis represents a phonological index for (r), where 100 would reflect a completely *r*-ful dialect and 0 would reflect a completely *r*-less one. The interview contexts that appear on the horizontal axis are designed to elicit increasingly careful, standardized speech. This figure shows that as the formality of the context increases, from casual speech through minimal pairs, so does the production of rhotic speech across all social groups. No group is categorically *r*-ful or *r*-less, and all groups exhibit a finely grained pattern of linguistic behavior that indicates consciousness of the *r*-ful form as the prestigeful target. On the basis of their collective orientation toward the same prestigeful targets across different variables—Labov studied (r), (th), and (-ing)—these randomly sampled New Yorkers could be classified as a single speech community. In the production arena, social differences are shown in patterns of variation. In the perceptual arena, social inferences are drawn from the architecture of variation.

Labov's (1966, 1972) study precipitated a scientific paradigm shift in the study of language and society. Since then, much sociolinguistic work has been carried out using the methodology of the *sociolinguistic interview*, a structured oral interview protocol that was originally designed to be administered to a large, randomly sampled, stratified urban population

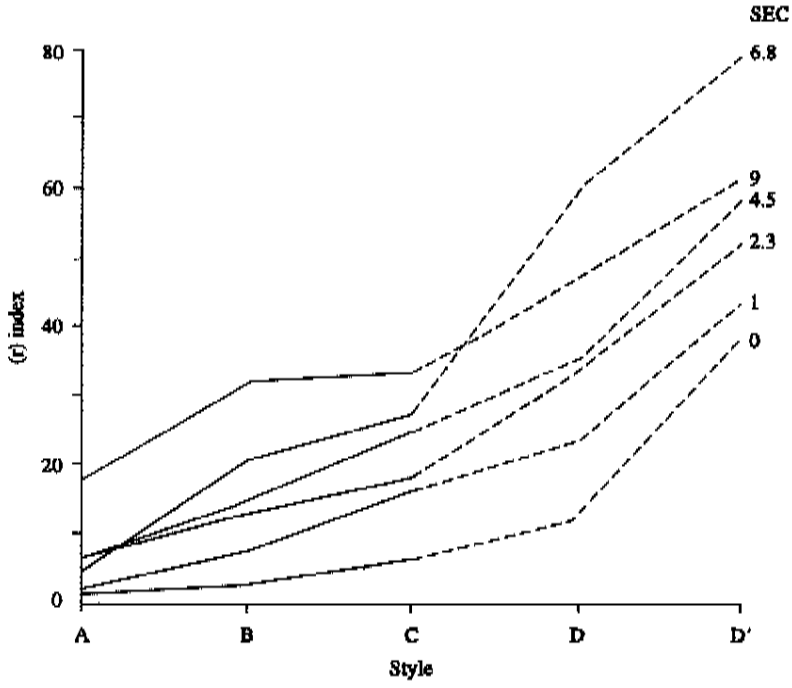

Figure 4.1
Class stratification of a linguistic variable in the process of change: (r) in *guard, car, beer, beard, board,* and so on. SEC (socioeconomic class) scale: 0–1, lower class; 2–4, working class; 5–6, 7–8, lower middle class; 9, upper middle class. A, casual speech; B, careful speech; C, reading style; D, word lists; D′, minimal pairs. (From Labov 1972, 114.)

of the sort studied by sociologists. Indeed, Labov's innovative interview method was first undertaken as part of a sociological survey of New York City. Soon thereafter, in the 1960s and 1970s, large-scale, quantitative studies began in other urban areas. Such studies aimed to model different strata of speech communities by including large numbers of speakers, varying with respect to age, ethnicity, socioeconomic status, and gender. Modern sociolinguistics is firmly grounded in the belief that language change is propelled by social variation, where innovative speakers push the envelope of preexisting changes, simultaneously abstracting from and constrained by structural linguistic factors. Linguistic facts that appear synchronically categorical—the lack of grammatical gender agreement in

English, for instance—appear from a diachronic perspective as the endpoint of a change that has been carried through to completion. Much of the motivation for presenting data on age-graded, style-based, or gendered stratification is to support claims of changes in progress. Indeed, the old notion of "free variation" has been entirely replaced in sociolinguistics by the notion of a change in progress, where the variable in question is assumed to be part of a system in flux, and the task of the sociolinguist is to identify, from a naturalistic sample, which is the conservative usage, which is (are) the innovative usage(s), who the innovators are, and what structural constraints they face.

To date, hundreds of urban studies in the United States and around the world have applied some version of the sociolinguistic interview methodology (though it is not without its problems—see Wolfson 1976; Briggs 1986), eliciting informal speech by asking interviewees what their childhood games were like, whether they have ever come close to death, and what kinds of changes they have experienced in their lifetimes (for a detailed explanation of the design of a sociolinguistic interview, see Feagin 2001). This method has proved remarkably productive and has served in creating finely stratified models of the speech of urban populations. This particular area of inquiry has come to be called *urban dialectology*, and here we cite but a few recent examples: Silva-Corvalán 1989 for Santiago, Chile; Thibault and Daveluy 1989, Thibault and Sankoff 1993 for Montreal, Canada; Tagliamonte 1999 for York, U.K.; Kontra and Váradi 1997 for Budapest, Hungary; Lennig 1978 for Paris, France; Trudgill 1974, 1988 for Norwich, U.K.; Horvath 1985 for Sydney, Australia; Rickford 1986 for Guyana; Haeri 1998 for Cairo, Egypt; and Labov, Ash, and Boberg, in press, for 145 cities in the United States alone, where sociolinguistic interview methodology and minimal pair elicitation have been combined to produce *The Atlas of North American English*.

4.2.3 Beginning with Frequency

Some of the first studies of language variation were done on sociolinguistic interview corpora, using frequency-based information to locate patterning in the production of linguistic variables. For instance, Wolfram (1974, 202) investigated linguistic contact among African-American and Puerto Rican speakers in New York City by examining their rates of monophthongization of /ay/. Monophthongization of /ay/ is understood

Table 4.1
Percentages of monophthongized /ay/ tokens in the speech of African-American speakers (AA), Puerto Rican speakers with extensive African-American contacts (PR/AA), and Puerto Rican speakers with limited African-American contacts (PR). (Adapted from Wolfram 1974.)

	AA	PR/AA	PR
No./Total	190/247	104/148	261/657
% monophthongized	76.9	70.3	39.7

to be an African-American English feature typically not present in Euro-American dialects in the northern United States such as that of New York City. Wolfram hypothesized that linguistic influence from African-Americans was the source of greater frequencies of monophthongization among Puerto Rican speakers with extensive contacts in the African-American community, as compared to those with limited contacts (see table 4.1). Although this brief example does not fully portray the complexity of Wolfram's findings, we will borrow it to help illustrate two extended points, one sociolinguistic-methodological and one mathematical.

First, in appealing to social explanations for the patterning of linguistic data, and to ensure their validity and replicability, students of variation begin by thoroughly investigating the social categories extant in a given community. Often this takes the form of prolonged ethnographic, participant observation fieldwork within the community in question. This particular feature of investigative inquiry minimizes the observer's paradox and creates a number of close connections between sociolinguistics and qualitative social sciences such as anthropology. In this case, Wolfram based his categorization on participant observation in addition to a follow-up interview designed to probe aspects of social contact between the African-American and Puerto Rican communities. Note that his categories go beyond census-based "ethnic" categories, instead reflecting associative groups in the community.

Second, the reasons behind quantitative variationists' shift in the direction of probabilistic approaches are also apparent in this example. Looking at the distribution of the variants in table 4.1 is not enough, for instance, to determine the comparability of the distribution of linguistic contextual factors in the interviews of different associative groups, or whether the contributions by subvariants within the variables are compa-

rable (Wolfram 1991, 25). Consider as examples of possible disparities two imaginary conditions: (1) that the distribution in the above case could be the result of a particularly frequent discourse marker that carries the monophthongized realization of the variable in question (such as *like* [la:k]); and (2) that such a marker is unevenly distributed in the speech community, with one of the groups using it much more frequently than the others. In such a case, we would have an irrecoverable distributional anomaly in the data, and the comparison of marginals (raw percentages) would be misleading. Providing frequency counts for each particular phonological context runs into a similar problem, since there are different numbers of total tokens in each group, and contexts before /k/ would be overrepresented in one group versus the other, causing a similar skew in those data.

And yet the following questions remain: Is the skew resulting from unevenly distributed linguistic contexts an artifact of the data collection method? Why do sociolinguistic data require collection methods different from those used in collecting other linguistic data? Couldn't all the distributional anomalies be easily avoided if the researcher controlled contexts and used laboratory elicitation? Part of the challenge of sociolinguistics is to take up the Malinowskian question introduced at the beginning of this chapter: shall we study language as a static entity, as it may occur word by word in isolation, or shall we study it as it unfolds in vivo, minimizing the effects of the laboratory and of the interviewer as much as possible?

The construction of a sample of naturally occurring speech is a different enterprise from the construction of a random sample in a demographic study, or of an experimental paradigm that can control exact numbers of presentations of stimuli, repetitions, ordering of contexts, and so on. Sociolinguistic data differ from census or experimental psychology data in that it is usually impossible to predict how often the relevant phenomenon will occur in the flow of naturally occurring conversation. Contributions to numerical skew and unreliability of pure proportional information and frequency counts may include the following:

1. *Unevenly populated speaker categories.* These may emerge because of distributional facts about the subject population, including rates of response in a door-to-door interview situation, or nature and number of participants in a naturalistic speech activity. Investigating a talk show situation such as *The Oprah Winfrey Show*, with a female talk show host

and a preponderance of female guests, easily illustrates such difficulties. These demographic difficulties as well as the time-intensiveness of transcription lead researchers to rely on a small sample size for speakers and to concentrate on collecting relatively long speech samples from each speaker, the standard length of a sociolinguistic interview being at least an hour.

2. *Widely disparate frequency of forms.* Certain variants of the variable in question may be possible but rare in naturalistic discourse. For example, Matsuda's (1993) study of analogical leveling found that some of the target variants of the potential forms of vowel-stem verbs seldom occurred in Tokyo Japanese discourse, with a frequency of four or five tokens per 90-minute interview. By its very design, the sociolinguistic interview is structured but not controlled, and additional methods may have to be devised (Matsuda's solution was to deploy ingeniously worded questions designed to elicit the elusive constructions (1993, 7)).

3. *High proportion of empty cells.* This is an extension of point 2, but often a mathematically fatal condition for certain kinds of statistical models (i.e., analysis of variance, chi-square) that demand controlled data. For example, phrases that appear to be possible in the combinatorics of generative syntax may be pragmatically restricted or may simply be unattested in the data set.

These factors contribute to the poor fit of sociolinguistic data to summary statistics such as percentages, and to analyses such as sum-of-squares approximations, setting the stage for multivariate probabilistic methods.

4.3 Incorporating Probability into Sociolinguistics

4.3.1 What Is/Was a Variable Rule?

Shortly following his first sociolinguistic studies of New York City, Labov (1969) proposed the *variable rule*. Working within the rule-based framework used in Chomsky and Halle's (1968) *The Sound Pattern of English*, Labov introduced the variable rule by distinguishing it from the categorical rule and

> associat[ing] with each variable rule a specific quantity ϕ which denotes the proportion of cases in which the rule applies as part of the structure of the rule itself. This proportion is the ratio of cases in which the rule actually does apply to the total population of utterances in which the rule can possibly apply, as defined by the specified environment. The quantity ϕ [in a variable rule] thus ranges between 0 and 1; for all categorical rules ... it follows that $\phi = 1$. (1969, 738)

This quantitative extension of the categorical rule framework was followed by a mathematical model developed by Cedergren and Sankoff (1974) and Sankoff and Labov (1979).

A new family of notational conventions accompanied the positing of this new theoretical possibility. One of the best-studied variables in sociolinguistics is word-final *-t/-d* deletion (e.g., [wes] for *west*), a common process that displays morphophonological, class-stratified variability in all English dialects. Variable rules soon ceased to be written with specific frequencies, because depending on a speaker's level of formality or social class the researcher would get differing frequency information, though the ordering and strength of constraints was similar for speakers in the same speech community (Fasold 1991). Thus, the constraints were assigned Greek alphabet letters in order of strength (α being the strongest). The following variable rule describes *-t/-d* deletion in Wolfram's data (Fasold 1991, 4; based on Wolfram 1974):

(1) $[\mathrm{d}] \rightarrow \langle \varnothing \rangle / \langle [-\gamma \mathrm{stress}] \rangle \langle -\beta \# \rangle \_\_\_\_ \# \# \langle -\alpha \mathrm{V} \rangle$

This rule states that word-final [d] optionally deletes when it is (1) in an unstressed syllable, (2) not a suffix, or (3) not followed by a vowel. Deletion is most likely when condition (3), the strongest constraint, is met.

Ordering the constraints is helpful, but it cannot fully describe which choice in the set will be used by a speaker as a member of a particular group. A probabilistic model can be derived to evaluate the contributing influences of each variable constraint. The VARBRUL family of statistical programs was originally developed by Rousseau and Sankoff (1978a) specifically to deal with the quantitative modeling of sociolinguistic data displaying the complexities described above. It is important to keep in mind the distinction between the variable rule theoretical framework for understanding sociolinguistic variation and the VARBRUL family of statistical programs, which is still used despite relative agnosticism by practitioners about what it actually models (Fasold 1991).

4.3.2 A Variable Rule Is Not a Generative Rule

The theoretical proposal of variable rules was immediately viewed with skepticism by generative grammarians and castigated in a series of articles, notable among which is Kay and McDaniel 1979. Here we examine the nature of this debate and its implications for underlying cognitive structures.

Although Labov, Cedergren, and Sankoff did not see the introduction of variable rules as a major departure from the concept of rules in linguistic theory, Kay and McDaniel argued that the variable rule was in fact such a radical departure that "it leads to a conceptual muddle in so far as its proponents think they are working within the generative framework" (1979, 152). To illustrate, Kay and McDaniel borrowed Chomsky's hypothetical context-sensitive rules for a simple natural language. Here rule (2b) is optional:

(2) a. $S \rightarrow ab$
b. $ab \rightarrow aSb$

These rules generate the set of all strings in a language where n instances of a are followed by n instances of b, as in $\{ab, aabb, aaabbb, \ldots\}$. Within this framework, there are different kinds of rules: obligatory rules like (2a), and optional rules like (2b) that specify more than one possibility in the derivation and allow for the generation of infinite sets of sentences with fixed rules. In terms of the hypothetical language above, a third optional context-sensitive rule might be posited, yielding strings such as $\{acbb, aacbbb, \ldots\}$:

(3) $a \rightarrow c \,/\, \_\_\_\, b$

This rule is already extremely close to being a variable rule in the sense introduced by Labov (1969). The only difference is that in addition to having contextual information, a variable rule has frequency information, and where (3) can be stated as "Realize a as c in the context before b sometimes," a variable rule might be stated as "Realize a as c in the context before b 69% of the time, when conditioned by the following variables ..." Kay and McDaniel argued that the leap from "sometimes" to a specific frequency is unwarranted, since "[t]he frequency with which a sentence is produced as an utterance (token) is completely irrelevant. Hence a 'rule' which is concerned with predicting token frequencies is not a rule of (generative) grammar" (1979, 153). Kay and McDaniel noted with alarm the break and incompatibility between the categorical nature of rules in closed, discrete, deductive-inferential systems and the gradient quality of the new variable rules, based on open-ended, continuous, and inductive-inferential systems (Romaine 1985; Givón 1979). But what are the different cognitive implications in these two representational statements?

Sankoff argued that "the formal models of grammatical theory have discrete structures of an algebraic, algorithmic and/or logical nature"

(1985, 75), allowing speakers to make a choice between two or more equivalencies (e.g., allophones) that might carry the same denotation. He continued, "By allowing a degree of randomness into the choice between such alternates, the grammatical formalisms are converted into probabilistic models of linguistic performance." Here, Romaine argued, is precisely where the chasm lies: generative grammars "do not generate true sentences or actual utterances, which are then checked against some corpus; they generate *correct* sentences.... In the most general terms, this type of grammar is a set of devices which check derivations for well-formedness" (1985, 59). Much like the laws of abstract algebra or subatomic physics, which cannot be tested against a corpus, so the aim of linguistic grammars is not to compare their output to naturalistic speech. Romaine further argued that if one were to truly extend the generative framework, a central characteristic of a sociolinguistic grammar would have to be sociolinguistic well-formedness. This sensitivity to social context is already about utterances in the world, and by its very violation of the principles of abstract derivation described above, it fatally fails to conform to the notion of what is meant as the object of description of a generative grammar.

Sankoff did not see variable rules as claiming a particular type of ontological status for the surface output they describe (Sankoff 1988), and yet Labov stated, "We can say that the kinds of solutions offered to problems such as consonant cluster simplification, copula deletion, and negative concord represent abstract relations of linguistic elements that are deeply embedded in the data. It is reasonable to suppose that they are more than the constructions of the analyst, they are the properties of language itself" (1972, 259). This does not necessarily imply that Labov believed in exact isomorphism between models and the phenomena described by the models, as Romaine suggested (1985, 65), but it does point to the possibility of understanding variable rules in two different ways: as a building block in a progressively more exact description of how humans cognitively organize language (Labov), or simply as a statistical "display tool" (Fasold 1991), which sociolinguists may use to discern the various influences in their data.

While during the 1970s much of the debate over variable rules revolved around challenges from generative theoreticians and increasing refinements in the mathematical model, urban dialectology scholarship from the 1980s onward split in two directions: one that adopted variable rules as a modus operandi and applied them in different sociolinguistic contexts

and to larger linguistic domains such as syntax (Weiner and Labov 1983; Rickford et al. 1995) and discourse (Vincent 1991); and one that challenged the use of variable rules altogether because of the perceived lack of a coherent stance on the nature of representation (Gazdar 1976; Sterelny 1983), or over the issue of whether percentages can be part of a speaker's knowledge of the language (Bickerton 1971; Butters 1971, 1972, later reversed in Butters 1990). Other challenges have arisen with the charge that because of their reliance on aggregate data, variable rules obscure information about individual performance (Itkonen 1983; Bickerton 1971; for a refutation see Guy 1980). Especially as generative linguists have moved away from rule-based frameworks and toward constraint-based frameworks like Optimality Theory and the Minimalist Program, most sociolinguists have been less inclined to make statements about the psychological reality of variable rules (Fasold 1991).

Fasold (1991, 10) observes that variable rules are designed to make objectivist predictions about the frequencies with which certain rules would apply under certain contextual conditions. However, we must also consider possible subjectivist probabilistic interpretations—choice models—of variable rules such as that espoused by van Hout (1984).

4.3.3 The VARBRUL Program

As a family of computer programs developed specifically to deal with the data of sociolinguistic variation, the VARBRUL programs are similar to logistic regression models. Practitioners working within the VARBRUL framework use the criterion of maximum likelihood estimation for determining how well a model with a given set of factors fits the data. The full details of the mathematical development of VARBRUL and its relationship to the variable rule framework appear in Cedergren and Sankoff 1974; Rousseau and Sankoff 1978a,b; Sankoff 1985, 1988; Sankoff and Labov 1979 (a reply to Kay and McDaniel 1979); and Rousseau 1989. Detailed instructions for employing the software are available in Young and Bayley 1996.

Binary logistic regression is also available in most modern statistics packages. It either goes by a name such as "logistic regression" (e.g., LOGISTIC in SAS, or Binary Logistic in SPSS) or can be implemented within a generalized linear model (e.g., GENMOD in SAS, or glm in S-Plus), by selecting a link function of "logit" and/or distribution of "binomial." One difference between VARBRUL and commercially available alternatives is the form of reporting of the coefficients, or

"weights," assigned to the selected independent variables. VARBRUL reports weights as probabilities, whereas other programs report them in logit form (i.e., as natural log of an odds). VARBRUL probabilities range between 0 and 1, with values below .5 indicating a disfavoring effect and values above .5 indicating a favoring effect. Corresponding logit values range between negative infinity and positive infinity, and when p is .5, the logit is 0. While no upper or lower bound exists for the logit, it is undefined when p equals exactly 1 or 0 (see discussion in Knoke and Bohrnstedt 1994, 334). Probability weights can be transformed into logit values by taking the log odds; that is, logit $= \log_e(p'/(1-p'))$. For further discussion of the logit function, see Manning, this volume, and Zuraw, this volume.

The formulas for the logistic or generalized linear model of VARBRUL in use today are as follows. Formula (1) is the generalized linear model:

$$\log\left(\frac{p}{1-p}\right) = w_0 + w_1 + w_2 + \cdots + w_n, \tag{1}$$

where w_0 is an input weight and $w_1 \ldots w_n$ are contextual factor weights. $\text{Log}(p/(1-p))$ is the logit function, while *log* stands for the natural logarithm (with base e).

For each n, w_n is equivalent to $\log(p_n/(1-p_n))$. Thus, (1) is equivalent to

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{p_0}{1-p_0}\right) + \log\left(\frac{p_1}{1-p_1}\right) + \log\left(\frac{p_2}{1-p_2}\right) + \cdots + \log\left(\frac{p_n}{1-p_n}\right), \tag{2}$$

where p_0 is an input probability and $p_1 \ldots p_n$ are contextual probabilities.

And since $\log xy = \log x + \log y$, (2) is also equivalent to (3), one of the most currently used multiplicative equivalents of (1):

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{p_0}{1-p_0} * \frac{p_1}{1-p_1} * \frac{p_2}{1-p_2} * \cdots * \frac{p_n}{1-p_n}\right). \tag{3}$$

VARBRUL estimates the contextual factor probabilities by combining the input probability (p_0, the likelihood that this variable "rule" may apply in the overall data set, regardless of any contextual influences) with the specific factor weights for all the factors included in the model.

Using a technique based on the dynamic clustering of Diday and colleagues (Sankoff 1985; Bochi et al. 1980), Rousseau (1978, 1989) further developed the log likelihood test, a procedure that tests whether a constraint has an effect that is significantly different from another in its constraint family. This test partitions the data into subsets and compares the difference between the log likelihoods of the subsets, comparing them to an analysis of the data without any partitions. From this test, it is possible to arrive at the optimal likelihood analysis as well as the optimal number of factors within each factor group.

4.4 Stylin' Oprah: A Case Study Exercise in Probabilistic Sociolinguistics

This section will illustrate the use of the VARBRUL program with an extended example drawn from our work on the speech of the American daytime TV talk show host Oprah Winfrey. We will begin with a description of the larger project and then discuss the application of the VARBRUL program to our data.

4.4.1 Data and Analysis

Our work attempts to describe variation in the speech of Oprah Winfrey. Style shifting (intraspeaker variation) in Winfrey's speech has been observed by other analysts and has been characterized as a device to appeal to a cross-section of viewers; most analyses in the literature have centered on topic and lexical choice (Lippi-Green 1997; Peck 1994).

Winfrey herself is originally from Kosciusko, Mississippi. She spent all of her language acquisition years in the U.S. South, attending high school and college (and beginning her broadcasting career at the age of 19) in Nashville, Tennessee. She later moved to Baltimore and then to Chicago where she currently hosts her show. We may then expect that in her speech she would draw from two overlapping repertoires: regional African-American English phonological features of the U.S. South and the supraregional speech variety that is normative in commercial broadcasting.

We suspected context-dependent style shifting at the sociophonetic level in Winfrey's speech and have thus far analyzed some early results on monophthongization of /ay/ (Hay, Jannedy, and Mendoza-Denton 1999; Hay, Mendoza-Denton, and Jannedy 2000). We call the phenomenon *monophthongization* simply for the sake of convenience. It is not our intent here to investigate which variant is underlying, the monophthongal

or the diphthongal one, but merely to posit that Winfrey's speech does vary, and that it does so in a patterned way. We do not assume an abstract representation for the phoneme /ay/; rather, we assume a distribution that encompasses the range from fully monophthongized to fully diphthongized forms.

Style shifting has been shown to be sensitive to many elements of the speech situation, including addressees, topics, referees, and even overhearers (Mendoza-Denton 1999). Rickford and McNair-Knox (1994) found that syntactic and phonological features of African-American Vernacular English covaried in the speech of an African-American teenage girl (Foxy) along two axes: Foxy's speech changed depending on whether her interlocutor was African-American or European-American, and depending on whether the topic was school-related or non-school-related (friendships and recreation). A similar result suggesting a strong unity of "styles" correlating with topics was found by the California Style Collective (1993), who looked at sociophonetic, prosodic, and discourse-marking features and their co-occurrence patterns in the speech of a Euro-American California teenager, nicknamed Trendy. Trendy's index of innovative features, like Foxy's, correlated with school topics, and even with subtopics, such as descriptions of individual groups of people within the social landscape of her school.

In our study, we have isolated samples of *The Oprah Winfrey Show* where Winfrey is talking into the camera or to a television studio audience, without a specific interlocutor. The lack of a specific addressee is crucial: this is the closest we can come in this naturally occurring situation to controlling for the effects of specific interlocutors. Concentrating on the absent persons to whom Winfrey refers in the various segments (who happen to be both topics and referees in this case) allows us to code the segments "about" a referee under a single code and to include the characteristics of these referees as our independent variables.

Most of the segments we coded were short passages describing a particular guest or brief announcements of upcoming shows. For instance, the following transcribed segment, all coded as keyed to the referee "Tina Turner," includes five examples of /ay/ (*my*, *wildest*, *Friday*, *night*, *trying*):

> But let me tell you about tomorrow's show. Tina Turner, we're following Tina around the country, Tina Turner made one of *my wildest* dreams come true, and you're gonna get to see it tomorrow, that's *Friday*. Actually last *night*, we were onstage dancing with Tina Turner. There's a brief look at our rehearsal: that's me, *trying* to keep in step with Miss Tina, you'll see that on tomorrow's show, it's great fun. (*The Oprah Winfrey Show*, May 2, 1997)

It is important here to note that our codings for individual referees are not strictly codings of referents but codings of global referee-as-topic. Thus, in this instance, the coding of the vowel in the first person pronoun *my* is keyed to the referee "Tina Turner," on the basis of prior findings about the importance of topics in the organization of variation (Rickford and McNair-Knox 1994).

A probabilistic model of sociophonetic-level behavior seeks to understand each instance of dependent variable /ay/ as a decision point for the speaker. Following the analogy of Preston (1991), the speaker must decide how to flip the variable coin: whether to pronounce the phonological diphthong /ay/ with a diphthongal phonetic realization [ay], a monophthongal one [a:], or something in between. For the purposes of reporting these results, we will look at the monophthongal realization as the surface variant we are trying to model. We begin with an input weight of .32 for the data set from this speaker (the likelihood that the monophthongal variant will occur across all contexts in her speech), since the monophthongal variant occurs about 32% of the time. Various independent variables such as situational characteristics, variables in the linguistic context, or characteristics of the referee will weight each particular "coin toss" in one direction or another. We attempt to account for factors that may modify this input weight and affect monophthongal realization either by promoting it or inhibiting it. In investigating whether Winfrey will choose to monophthongize /ay/ (if indeed this process can be characterized as residing solely in the speaker's choice space), the question we mean to ask through the use of probabilistic methodology is: What possible social or linguistic factors, or their combination, influence this choice? Possible factors might be sociodemographic characteristics of the referee (in this case the African-American singer Tina Turner), the phonological and prosodic environments of the segment, or the frequency of the carrier lexical item. To test these questions, we coded the data with factors that include both the linguistic or "internal" and referee-sociological or "external" factors.

We coded 229 words containing /ay/ taken from discontinuous selections of approximately six hours of *The Oprah Winfrey Show*, from segments that aired in the 1996–97 season. We examined tokens by means of both auditory and acoustic criteria. Two phonetically trained listeners performed an auditory analysis of the data: a token was coded as monophthongized if and only if the listeners agreed on the classification. To provide acoustic verification of the auditory analysis, the vowel quality

was coded on the basis of spectrographic displays: each token in the data set was labeled either as a monophthong or as a diphthong from wide-band spectrograms. Although monophthongization of /ay/ is a continuous phonetic phenomenon, for the purpose of data entry into the VARBRUL program it must be treated as discrete: preferably as a binary variable, ternary variables being possible but necessitating collapse into the most predictive binary set. This limitation is one of the disadvantages of using VARBRUL analysis when treating continuous variables. Its implications are considerable and will be discussed at length in the next sections, where we compare VARBRUL analysis with other possible analyses.

We were able to distinguish three auditory possibilities for the realization of /ay/: fully diphthongized, fully monophthongized, and somewhere in between. Statistical analyses were carried out for two possible groupings of the tokens in the data set: one that considered only the fully monophthongal tokens as monophthongs, and one that considered both the slightly monophthongal and the fully monophthongal tokens in one category. According to these analyses, the most predictive and consistent results emerged with the latter grouping. Of the 229 tokens of /ay/ in our sample, 32% (74/229) were monophthongized according to the more inclusive definition, and 68% (155/229) were diphthongs. Since the diphthongal realization of /ay/ is normative in the standard language of the media, it is noteworthy that one-third of the tokens were monophthongal.

All the factor groups initially tested in this analysis are listed in table 4.2; statistically significant results, with raw frequencies and probability weights, are reported in table 4.3.

4.4.2 Explanation of Factor Groups and Results

The data were analyzed using Goldvarb Version 2.0 (Rand and Sankoff 1990), a variable rule program for the Macintosh computer. Both the application and its documentation are available online at ⟨http://www.CRM.UMontreal.CA/~sankoff/GoldVarb_Eng.html⟩.

Widely accepted by sociolinguists, the VARBRUL family of programs of which Goldvarb is a member utilizes the maximum likelihood estimate (Sankoff 1988) discussed above. Goldvarb computes probability weights that are expressed as likelihoods, with a probability weight of .5 neither favoring nor disfavoring application of the process in question. Probability weights between .5 and 1 favor the process more strongly the closer they are to the asymptotic 1, while probability weights between .5 and 0

Table 4.2
Variables, factor groups, and factors tested in study of monophthongization in the speech of Oprah Winfrey

Variable status	Factor groups	Factors
Dependent variable	monophthongal vs. diphthongal /ay/	diphthongized slight monophthongization full monophthongization
Independent variables (linguistic/ internal)	preceding phonetic context	voiced obstruents voiceless obstruents nasals liquids vowels/glides
	following phonetic context	voiced obstruents voiceless obstruents nasals liquids vowels/glides
	word class	open closed
	frequency in corpus	infrequent = occurring < 5 times in corpus frequent = occurring > 5 times in corpus
	log-converted CELEX frequency	unattested < log 2 between log 2 and log 4 between log 4 and log 6 between log 6 and log 8 between log 8 and log 10 between log 10 and log 12 < log 12
Independent variables (social/external)	referee gender	male female indeterminate or inanimate
	referee ethnicity	African-American zero referee non-African-American
	individual referee	18 individual referees (see appendix) were given separate codes; "other" category was also used
Variable interactions (social/linguistic)	ethnicity and frequency	African-American infrequent (< log 10) African-American frequent non-African-American infrequent non-African-American frequent zero infrequent zero frequent

disfavor the application of the process more strongly the closer they are to asymptotic 0.

VARBRUL analysis makes the mathematical assumption of an ideal data set with crosscutting factor effects but without significant interactions, where all the factors are independent of one another (Sankoff 1988, 4–19). However, certain factors in this data set are extremely likely to display collinearity. In practice, then, many factors (like word class and raw frequency), being highly correlated, could not appropriately be run together. As a result, only factors that could be assumed to be fairly independent of each other were run together. It is widely believed that social factors may show a high degree of correlation (Bayley 2001), but researchers think that it is relatively rare to find correlations across internal and external variables. Labov (2001, 84) states:

> A full assessment of the effects of intersecting social parameters, and a complete account of sociolinguistic structure, is only possible with multivariate analysis. A multivariate approach was first introduced into sociolinguistic studies in the form of the variable rule program (Rand and Sankoff 1990). It was motivated not by the need to analyze external, social factors, but rather to deal with the language-internal configuration of internal, linguistic constraints on variation (Cedergren and Sankoff 1974). The basic fact about internal factors that the variable rule program continually displays is that they operate independently of each other (Sankoff and Labov 1979). However it was realized from the outset that social factors are typically not independent. Though it is convenient and useful to incorporate external and internal factors in the same analysis, a considerable amount of information can be lost in the typical VARBRUL analysis of speech communities.

Including both internal and external factors is crucial to our data, however, since we found an interaction between lexical frequency calculated on the CELEX database (Baayen et al. 1995), presumably a purely linguistic variable, and ethnicity of referee, a social variable. The results presented in table 4.3 are explained by factor group below.

4.4.2.1 Preceding Phonetic Context and Following Phonetic Context We coded the immediately surrounding phonetic context of each /ay/ token within and across syllables and words, utilizing categories that have been shown in the literature to affect monophthongization in both African-American and Euro-American U.S. South populations.

Coding monophthongization according to a sonority hierarchy (Selkirk 1984) follows widely accepted methodology outlined by Hazen (2001). We included two other categories as well: vowel/glide and pause. Several studies (Thomas 1995; Schilling-Estes 1996; Wolfram, Hazen, and

Table 4.3
Raw frequencies and VARBRUL probability weights for significant categories in the analysis of monophthongal /ay/. Application = monophthongal /ay/; nonapplication = diphthongal /ay/. First run, with factor groups 1, 2, 3: input 0.323, log likelihood = −110.371, $p < .005$. Second run, with factor groups 1 and 4 only: input 0.263, log likelihood = −106.939, $p < .000$.

Factor group	Factors	Apps.	Non-apps.	Total *N*	% of total *N*	VARBRUL probability weight
1 Following segment	vowel/glide	9	35	44	19	0.799
	%	20	80			
	liquid	22	17	39	17	0.804
	%	56	44			
	nasal	13	35	48	21	0.436
	%	27	73			
	voiced obstruent	9	35	44	19	0.384
	%	20	80			
	voiceless obstruent	13	62	75	33	0.320
	%	17	83			
	pause	3	1	4	2	unreported owing to low token count
	%	75	25			
2 Ethnicity of referee	zero referee	19	15	34	15	0.700
	%	56	44			
	African-American referee	39	49	88	38	0.622
	%	44	56			
	non-African-American referee	16	91	107	47	0.336
	%	15	85			

3 CELEX frequencies						
3a Log value, 5-way split	< log 6	1	20	21	9	0.063
	%	5	95			
	log 6–log 8	17	33	50	22	0.478
	%	34	66			
	log 8–log 10	10	39	49	21	0.418
	%	20	80			
	log 10–log 12	17	36	53	23	0.596
	%	32	68			
	> log 12	29	27	56	24	0.734
	%	52	48			
3b Log value, binary split	infreq = < log 10	28	92	120	52	0.370
	%	23	77			
	freq = > log 10	46	63	109	48	0.642
	%	42	58			
4 Interactions	African-American infrequent	12	31	43	19	0.437
	%	28	72			
	African-American frequent	27	20	47	21	0.781
	%	57	43			
	non-African-American infrequent	6	58	64	28	0.177
	%	9	91			
	non-African-American frequent	10	31	41	18	0.576
	%	24	76			
	zero infrequent	9	4	13	6	0.783
	%	69	31			
	zero frequent	10	11	21	9	0.725
	%	48	52			
All totals	token count	74	155	229		
	% of data	32	68			

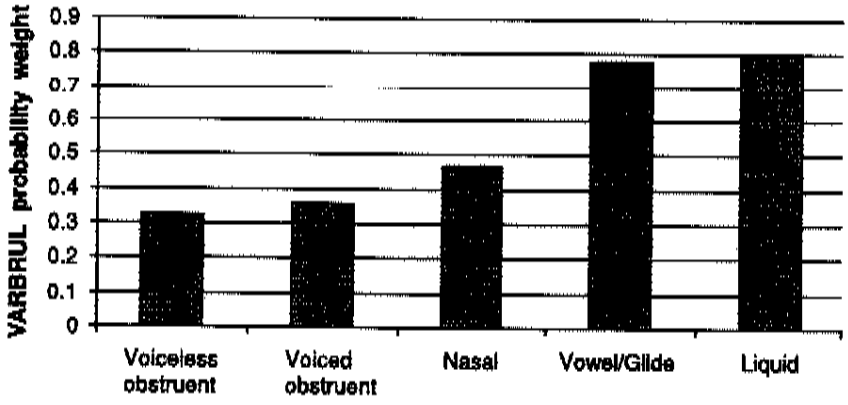

Figure 4.2
VARBRUL weights for following phonetic category as a predictor of monophthongization. Values above 0.5 favor monophthongization, values below 0.5 disfavor.

Schilling-Estes 1999) have shown that the expected descending order of following phonetic environments favoring monophthongization for *African-Americans* is liquids > nasals > voiced obstruents > voiceless obstruents. Our data for following phonetic context fit the expected pattern, with probability weights as follows (see also figure 4.2): liquids .804 > vowel/glide .799 > nasal .436 > voiced obstruent .384 > voiceless obstruent .320. In this data set, following voiceless and voiced obstruent contexts heavily disfavored the expression of monophthongal /ay/, while a following nasal neither disfavored nor favored the process. Only liquid and vowel/glide following contexts strongly promoted the expression of monophthongized variants. Because Euro-American U.S. Southerners exhibit a different pattern, with high rates of monophthongization before voiced obstruents, we believe that Winfrey's use of /ay/ is indexical of variation in the African-American community.

Preceding phonetic context was not a significant predictor of variation in this data set and was discarded in the final analysis.

4.4.2.2 Word Class As with other reductive processes (Wright 1997), monophthongization may apply at different rates among words depending on their frequency. In an earlier analysis, we tested lexical frequency within the Oprah Winfrey corpus and found that it was highly correlated with monophthongization (Hay, Jannedy, and Mendoza-Denton 1999).

Frequency can be a difficult metric to use because it may be partly confounding a linguistic factor: whether a word belongs to a closed class or an open class. To test this confound, we coded open and closed classes separately from frequency. When run by itself as the only independent variable, word class is highly significant in predicting the patterning of our data. Open class words disfavored the monophthongization process with a probability weight of .397, while closed class words favored it with a weight of .643 (log likelihood = −137.896, $p < .001$). Although both word frequency and word class were significant on their own, the most predictive model of the data was found by using the log-converted CELEX frequency category (see section 4.4.2.4).

4.4.2.3 Raw Frequency in the Corpus One issue when trying to use lexical frequency as a predictive factor in the study of a naturally occurring sample is whether to use the word frequency of the sample itself or some independent metric of word frequency in the language as a whole (because the sample might not be representative of the speaker's overall repertoire). In our case, words that were very frequent in the sample were words like *style* (from a segment of *The Oprah Winfrey Show* called "The House of Style") and *wild* (the descriptor of Tina Turner's "Wildest Dreams" tour).

As a first step toward assessing the importance of frequency, we used the raw frequency within our corpus and divided the words into "frequent" (>5 occurrences in the sample) and "infrequent" (all other words). This distinction also yielded significant results: infrequent words disfavored monophthongization with a probability weight of .329, while frequent words slightly favored it with a weight of .589 (log likelihood = −138.474, $p < .001$). Although significant on its own, raw frequency in this corpus was overshadowed by the log-converted CELEX frequency, which contributed more substantially in fitting the model to the data.

4.4.2.4 Log-Converted CELEX Frequency Another frequency metric that we used was frequency in the English language according to the CELEX corpus. The CELEX database (Baayen et al. 1995) from the Max Planck Institute for Psycholinguistics in Nijmegen incorporates the 17.9-million token COBUILD/Birmingham corpus, and in addition represents more than 90,000 lemmas from dictionary entries. All the sources for CELEX are textual, about 15% coming from U.S. authors.

Despite the differences (oral vs. textual, U.S. vs. composite) between our raw frequency corpus and the CELEX corpus, CELEX codings were better able to account for variation in our data. This strongly suggests that the processes at work in the patterning of our data transcend these particular instances of *The Oprah Winfrey Show* and may well be operating in other contexts as well.

The CELEX ordinal frequency ranking for each token was converted to a log-based frequency code because there is good evidence that humans process frequency information in a logarithmic manner. That is, a frequency difference occurring among the lower frequencies carries more weight than a frequency difference of equal magnitude occurring among the higher frequencies. Since VARBRUL requires discrete independent variables, in order to input the data we created a five-way log value split that provided a near-perfect cline of influence in which the most frequent words (> log 12) strongly favored monophthongization (probability weight .734), while the least frequent words (< log 6) strongly disfavored it (probability weight .063) (see figure 4.3). A binary (median) log value division was also devised. Words of frequency < log 10 strongly disfavored monophthongization (probability weight .370), while words of frequency > log 10 favored it (probability weight .642) (see figure 4.4). We used the binary division to code for interactions between word frequency and ethnicity.

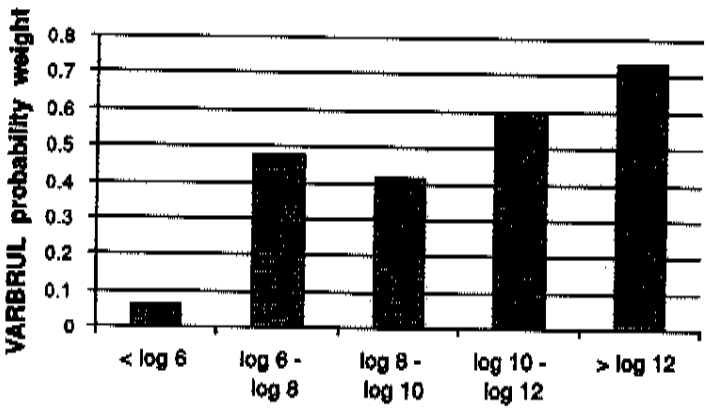

Figure 4.3
VARBRUL weights for lexical frequency as a predictor of monophthongization: results for log-converted CELEX frequency. Values above 0.5 favor monophthongization, values below 0.5 disfavor.

4.4.2.5 Individual Referee To investigate the possibility that Winfrey was treating each referee in an idiosyncratic or individualized way, and not according to gender or ethnicity, we also assigned segments codes that referred to people individually. Nineteen referee codes (including "other" for segments that were not about particular people) were used (the full list is given in the appendix). The "other" code was used primarily when Winfrey spoke straight into the camera without a specific addressee. These segments fulfilled our requirement that the speech have no specific interlocutor, and they included a segment called "Gratitude Moments," where Winfrey spoke about her favorite things, one where she spoke about her birthday, and one where she warned the audience about getting scammed (robbed). One of our initial hypotheses was that the individual codes would be important predictors of variation. However, it was not borne out in the VARBRUL results and was eliminated. In our later analysis using CART trees, the individual codes became an important factor.

4.4.2.6 Referee Gender So-called external or social variables that we coded in the corpus included the referee's gender. By itself, again, gender was significant, but not when included in a statistical run with any other factor. Codings for this factor included male, female, and "other," used for situations where the referent did not have a gender or where gender

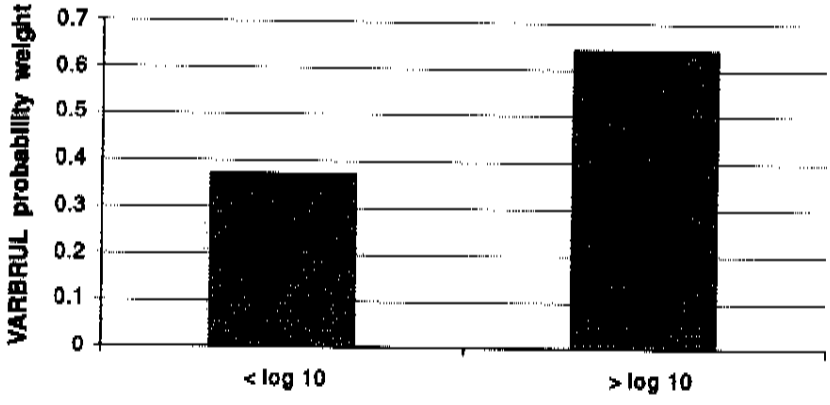

Figure 4.4
VARBRUL weights for lexical frequency as a predictor of monophthongization: results for log-converted CELEX frequency (cutoff at median). Values above 0.5 favor monophthongization, values below 0.5 disfavor.

could not be determined. Oddly enough, there were no statistically significant differences between rates of monophthongization for female and male referees (both of these were neutral, f: .468, m: .435), while the "other" category showed a markedly favoring effect, o: .730 (log likelihood = −139.252, $p < .01$).

4.4.2.7 Referee Ethnicity The ethnicity of the referee was the most important factor group (first selected in the Goldvarb step-up/step-down procedure) in modeling the monophthongization of /ay/. We coded referee ethnicity according to three categories: African-American referees (strongly favoring monophthongization; probability weight .622); non-African-American referees (strongly disfavoring; .336); and zero referee, which favored monophthongization more strongly (.7) than the other categories. These weights are shown in figure 4.5. However, it was also clear from our analysis that ethnicity of referee also interacted strongly with word frequency. And VARBRUL assumes that the different factors included in a single analysis act independently of one another.

4.4.2.8 Ethnicity and Frequency One solution to the problem of assuming factor group independence in VARBRUL is to create an inde-

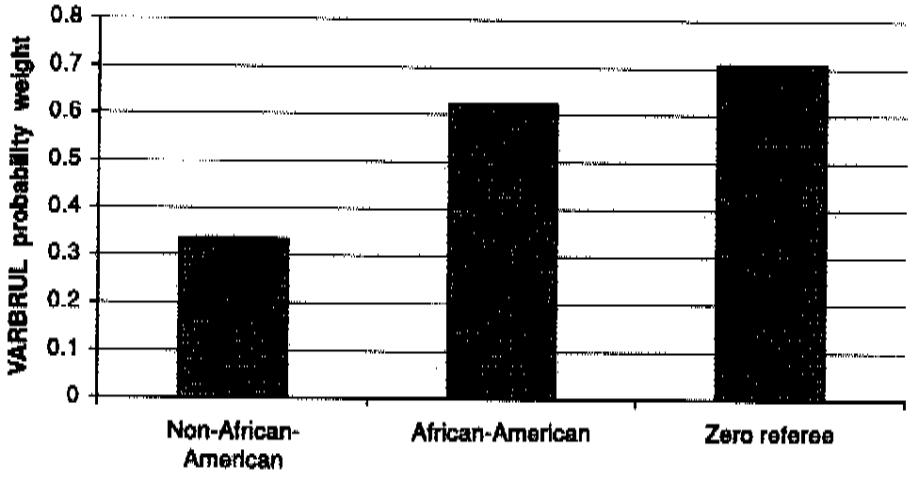

Figure 4.5
VARBRUL weights for ethnicity of referee as a predictor of monophthongization. Values above 0.5 favor monophthongization, values below 0.5 disfavor.

pendent factor group that combines the interacting factors into discrete possibilities in order to isolate their effects (Bayley 2001). We did this by creating an additional interaction factor group that combined the six possibilities resulting from the interaction of two frequency categories (frequent vs. infrequent words; i.e., > log 10 vs. < log 10 in the binary CELEX coding) and three ethnicity categories (African-American, non-African-American, and zero referee). Our results were most puzzling: they showed a significant interaction in what we had originally coded as two separate predictive factor groups. When combined, the ethnicity of referee/binary CELEX frequency factor group was intriguingly arranged thus (see also figure 4.6): [no ref, infrequent .783 > African-American, frequent .781 > no ref, infrequent .725 >] non-African-American, frequent .576 > African-American, infrequent .437 > non-African-American, infrequent .177. The bracketing around the first three factors indicates that according to the difference-in-log-likelihoods test (Rousseau 1989), these factors are not statistically significantly different from each other and should be collapsed. They are shown separately here for expository reasons.

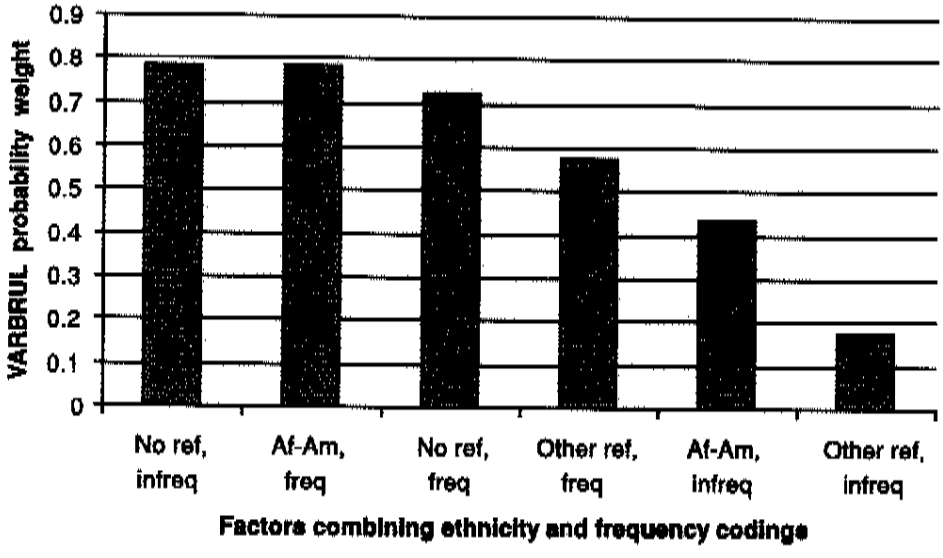

Figure 4.6
VARBRUL weights for the interaction between referee ethnicity and lexical frequency as a predictor of monophthongization. Values above 0.5 favor monophthongization, values below 0.5 disfavor.

Initially, we found this result—that either a frequent or an infrequent word with a zero referee was as likely to lead to monophthongization as a frequent word with an African-American referee—difficult to explain, especially since there is such a strong distinction in likelihood of monophthongization between African-American and non-African-American referees. What could this mean? Colleagues have suggested to us that zero referee might be Winfrey's baseline style, and that this might be close to a style used for African-American referees. And, as discussed below, much of the zero referee section included frequent self-reference, specifically using the words *I* and *my*. Of course, self-referring words are also frequent words, so it is difficult to disentangle the two effects and precisely identify the locus of the observed patterns. Because the two predictors are highly correlated with one another, one cannot simply include them both in a logistic regression, to see which one is the stronger predictor. The technique we have used assumes strict independence among the factors. In the next section, we explain how we set about investigating the self-reference effect. For now, we return to the interaction between ethnicity of referee and lexical frequency.

From the interaction of frequency and ethnicity of referee, given our understanding of frequent words as the carriers of style (Hay, Jannedy, and Mendoza-Denton 1999) we expected frequency to have a much bigger effect in speech relating to African-American referees (where frequent words should be prone to monophthongization for both stylistic and articulatory reasons) than in speech relating to non-African-American referees (for whom we expect some canceling out of the articulatory tendency to monophthongize frequent words, given that the stylistic setting favors diphthongs). In fact, our results show the opposite: word frequency has a much bigger effect for non-African-American referees than for African-American referees. Upon closer examination, we believe this result does not necessarily contradict our assumptions about frequency and style; rather, it reflects an asymptote for this type of variation. When referencing African-Americans and using frequent words, Winfrey reaches the limit of her range of variation. VARBRUL probability weights around .78 set the upper bound of monophthongization that can be found in Winfrey's speech. Essentially, there is a ceiling effect, indicating that in no speech situation will she go beyond her personal maximum of variation (even just doubling her rate for infrequent words with African-American referees would overshoot this asymptote).

Frequency thus has the biggest effect within the subset of the data that is not otherwise prone to monophthongization (non-African-American referees), while referee ethnicity has the biggest effect within the subset that is not otherwise prone to monophthongization (the infrequent words).

4.4.2.9 Self-Reference, Style, and the Use of Qualitative Analysis To disentangle the effects of lexical frequency and ethnicity, we inspected the show transcripts and found specialized discourse patterns in the use of the highly frequent words *I* and *my*. The segments coded as "zero referee" consisted largely of Winfrey self-disclosing to her audience. The segments "House of Style," "My Favorite Things," and "Oprah's Birthday" are frequently self-referring. This self-disclosure feature of Winfrey's television persona—and the genre of daytime talk shows in general—has received a great deal of attention from scholars (Shattuc 1997; Masciarotte 1991). In terms of our data, a style of conversational engagement through self-disclosure means that Winfrey talks about her own past encounters with the people to whom she refers, sharing her personal history in great detail. Guests she knows well elicit more self-reference, so that a short segment on Michael Jordan, with whom Winfrey has a famously close relationship, included 8 self-referring tokens out of 17 /ay/ tokens, or 47% of the tokens for that segment. The segment "My Favorite Things/Birthday Presents" included 23/35 or 65% self-referring tokens. A segment about Mia Farrow, by contrast, included only 1/20 or 5% self-referring tokens.

The use of highly frequent words as stylistic devices in the genre of talk show hosting may boost overall perceptual saliency of the variable and make it a good candidate for the display of speaker style. Other examples of highly frequent words used as iconic displays of speaker and group style can be found in Mendoza-Denton 1997 and California Style Collective 1993.

When included in a VARBRUL run with ethnicity and with following phonetic context only, self-referring words joined these factors in a set that best predicted variation in the data. Self-referring words correlated positively with monophthongization, exhibiting VARBRUL weights very similar to those of the zero referee category; non-self-referring words had probability weight .398, while self-referring words had weight .680 (log likelihood = -109.428, $p < .001$) (see figure 4.7).

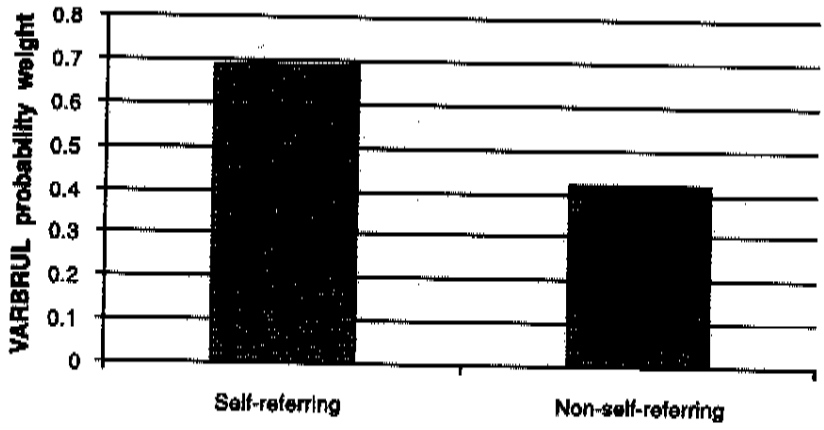

Figure 4.7
VARBRUL weights for self-reference as a predictor of monophthongization. Values above 0.5 favor monophthongization, values below 0.5 disfavor.

We believe both frequency and self-reference are playing a role in the aggregate data set. The observed frequency effects are spread throughout the whole frequency range (see table 4.3, where frequency effects were significant even when split into five frequency categories), and so they cannot be attributed only to self-reference. However, it is consistent with the observed discourse patterns to hypothesize that Winfrey's self-referring speech might be particularly prone to monophthongization—although, as explained above, because of the collinearity of the factors this is best investigated qualitatively. Precisely disentangling the relative contribution of frequency and self-reference would require a larger data set and remains a task for future work.

4.4.3 An Alternative to VARBRUL: Classification and Regression Trees (CART)

In this section, we briefly explore the patterns in our data further, using a different statistical approach.

The construction of classification trees is essentially a type of variable selection. Such trees are a valuable tool for exploratory data analysis and can handle missing values or empty cells with ease, tree construction being based on the cases that do not have missing values. Classification trees are an attractive method of data exploration because they handle interactions between variables automatically. They also have the advan-

tage of being completely nonparametric. No assumptions are made about the underlying distribution of the data. These features make them less powerful for detecting patterns in data, but fairly reliable in terms of the patterns found.

Classification trees do assume that the effect being modeled is organized into discrete factors. An analogous class of models, regression trees, deals with continuous data.

Foundational literature on classification and regression trees includes Morgan and Sonquist 1963, Morgan and Messenger 1973, and Breiman et al. 1984. A good practical guide for their implementation in S-Plus can be found in Venables and Ripley 1994.

A classification tree begins with the data to be analyzed and then attempts to split it into two groups (here, one that maximizes monophthongization, and one that minimizes it). Ideal splits minimize variation within categories and maximize variation across categories. All possible classifications of the independent variables are attempted. Tree construction works one step at a time, so once the first split is achieved, an optimal split is sought for each resultant node. The particular technique used here (that implemented in S-Plus/R) allows only binary splits. At any given node, the maximum reduction of deviance over all possible splits is used to identify the best split. This process continues until either the number of cases reaching each leaf is small or the leaf is sufficiently homogenous relative to the root node.

This process often grows a tree that overclassifies the data. That is, a tree may fit a particular data set extremely well, but may be unlikely to generalize if new data points are added to the analysis. A selection process can then be used (akin to the stepwise procedure used in multiple regression) to determine which divisions should appropriately be included in the model and which are best discarded—a process known as tree pruning (Breiman et al. 1984). There are a number of different methods for choosing where to prune the tree (i.e., for deciding which nodes can best be removed).

One method of tree pruning uses a process of cross-validation. The data set is divided into subsets, and separate trees are grown on the basis of each subset. The trees based on each subset of the data can then be compared with one another. As Venables and Ripley (1994, 44) explain, "Suppose we split the training set into 10 (roughly) equally sized parts. We can then use 9 to grow the tree and test it on the tenth. This can be done in 10 ways, and we can average the results." This process returns an

averaged deviance for trees of each possible size. In the analysis presented below, we used this cross-validation technique—pruning the tree to the smallest tree size with the minimum deviance. This represents a fairly conservative approach to tree building.

When we attempted to build a tree based on the monophthongization data, we allowed for the possible contribution of the following variables: the individual identity, ethnicity, and gender of the referee; the class and frequency of the word; the preceding and following phonetic environment. Of these, only two remained in the pruned tree: the identity of the individual and the following phonetic environment. Because each branch of the tree deals with successively smaller sets of data, a fairly large data set is required to establish the coexisting significance of a sizable number of contributing factors. The power of this technique is therefore slightly limited when dealing with small data sets—especially if these data sets display much variability.

The pruned tree is shown in figure 4.8. The first and most important split is between segments where Winfrey is talking about Tina Turner, Will Smith, Halle Berry, or no one in particular (group (c)), and all other segments. In the former four instances, she was much more likely to monophthongize /ay/ (60% of tokens) than in all others (17%).

These two nodes split further into two subcases. The left branch splits into two more sets of individuals: those who strongly discourage monophthongization (group (a): 2%) and those who are more likely to lead to monophthongization (group (b): 23%). Finally, among group (c) the classification algorithm detects a significant effect of the following environment: monophthongization is more likely preceding liquids and nasals than other phonological segments.

Other variables were included in the full tree (lexical frequency is the next factor to appear), but did not survive the pruning process. Because a classification tree looks for patterns in progressively smaller sets of data, we would likely need a much bigger data set than we currently have in order for it to reveal the full range of complexity in our data. Those factors that do survive the pruning process, however, are ones in which we can have extreme confidence.

The classification algorithm divides individuals into three groups. No African-American referee appears in group (a), 3 African-American referees appear in group (b) (3/8, 38%), and the individuals identified in group (c) are all African-American and are grouped together with the zero referee cases.

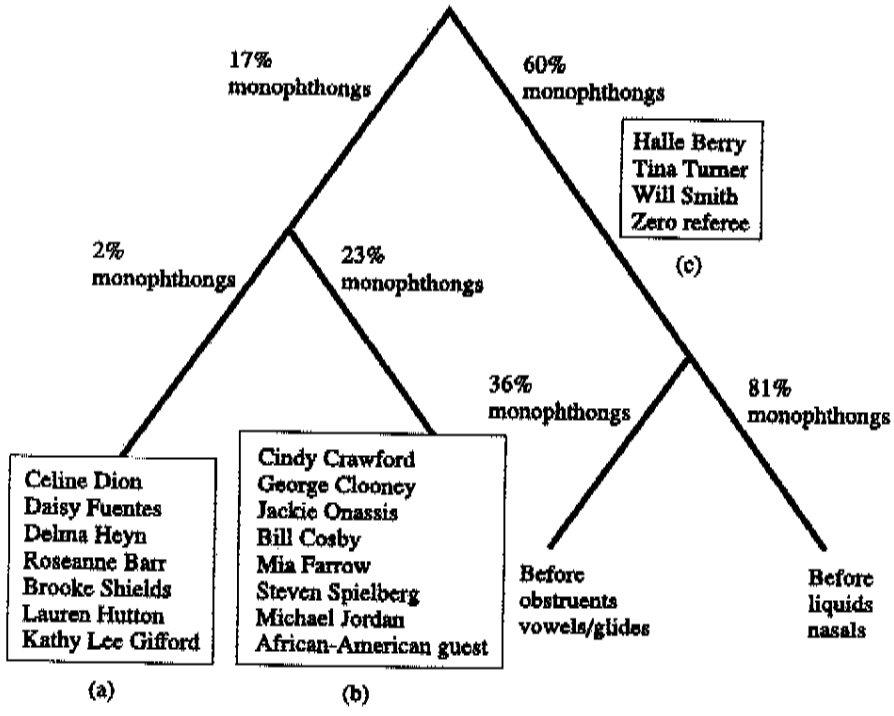

Figure 4.8
CART classification tree for monophthongization

This "individual identity" variable therefore echoes the effects of ethnicity, while imbuing it with an added level of subtlety. It could perhaps be seen as organizing people into groups according to the nature of Winfrey's involvement with them—ethnicity being one component of this (though probably not a binary component), and other dimensions of solidarity perhaps also playing a role. Because the classification tree is an excellent tool for revealing significant groupings in data, it can be used to reveal natural groupings of individuals, which (as inspection of the groups reveals) could not be replicated by any combination of standard social variables.

4.4.4 The Oprah Winfrey Data: Summary

4.4.4.1 Analysis Techniques Using a relatively small data set, we have shown how inferences can be derived from it in different ways. What we

hope to have demonstrated with this exercise is that different types of analysis can assist in the interpretation of results. In our case, we used two types of quantitative analysis (VARBRUL and CART) as well as qualitative analysis (looking at patterns of self-reference). Two basic results emerge unambiguously from our study: Winfrey's style shifting is partially conditioned by the ethnicity of the person she is referring to, and partially by the following phonetic environment.

Subtler nuances of the data—the role of lexical frequency, the presence of interaction effects, the emergence of natural groupings of individuals—are highlighted differently by the different statistical techniques we used.

Each technique has drawbacks. Since classification trees are local optimizers, once an initial split is made it is impossible to ask what the overall effect of a second factor is, given the first one. And in order to examine the effect of a large number of variables using a classification tree, a large data set is required. VARBRUL is not well equipped to easily explore possible interactions, nor is it equipped to deal with continuous dependent or independent variables, although these limitations can be overcome by the use of logistic regression in commercially available statistics packages.

The VARBRUL program effectively implements a binomial stepwise regression analysis. It models a binomial outcome, using discrete factors. In this sense, it is an appropriate tool to use in the formulation of variable rules as they were originally conceptualized—rules that predict which of *two discrete outcomes* will occur on the basis of *discrete factors*, such as the gender or ethnicity of the speaker (or in the case of our data, the referee) or the identity of the phoneme that precedes or follows. Continuous independent factors can be built into the model, by breaking them up into discrete groupings—a technique that imposes artificial category boundaries on the factor.

And yet monophthongization is not really discrete. Different degrees of monophthongization (or diphthongization) exist, and Winfrey exploits the full range of this continuum in her performance of style. Winfrey does not shift between discrete styles (an "African-American referee," a "non-African-American referee," and a "self-disclosure" style); rather, she seamlessly navigates a range of continuous stylistic dimensions, and the degree to which she employs monophthongization signals (together with many other variables) where she is positioned in stylistic space. Monophthongization is not discrete, ethnicity is not discrete, nor is lexical

frequency. And yet we impose boundaries on all of these in order to simplify our analysis and detect statistical patterns that will shed light on language and on society.

We believe one challenge for the future of probability theory in sociolinguistics is to move beyond the limitation of discrete categorization and to work toward understanding how gradient, continuous linguistic variables are conditioned by both categorical and continuous social and linguistic factors. Such analyses have begun to appear, notably Berdan's (1996) study of second language acquisition, where he used the logistic regression module in SPSS to model time as a continuous independent factor. Sudbury and Hay (in press) also model continuous independent factors (time and frequency) in their analysis of rhoticity and /r/-sandhi.

Modeling and understanding the combination of continuous and discrete factors in predicting gradient implementation of sociolinguistic variables will be a major challenge in the future of probability theory in sociolinguistics—one that will require an adjustment in the way data are collected and analyzed, and in the statistical techniques used to explore the patterns of variability within those data.

As illustrated in our Oprah Winfrey data analysis, if one of the predictor variables in the hypothesized model is continuous (such as lexical frequency or age), VARBRUL is unable to model it as a continuous predictor; instead, the researcher must break it up into a number of discrete sets. This does not tend to be a feature of more general implementations of logistic regression, which can unproblematically model continuous variables. Thus, as discussed by Berdan (1996) and Bayley (2001), the VARBRUL implementation may not be the most appropriate for data sets that involve one or more important continuous independent variables.

And while it is possible to encode interactions in VARBRUL by creating hybrid categories (see, e.g., Sankoff and Labov 1979, 204; also the example in section 4.4.2.8), this solution is not straightforward, and it requires that the researcher identify the possible interaction in advance. Other implementations of logistic regression tend to allow possible interaction effects to be explored in a more straightforward way. Sigley (2001) tested for the presence of interactions in seven previously reported data sets and found that about 26% of pairwise tests produced significant interactions. He argues that interaction effects are widespread and are

potentially just as important as main effects when modeling (socio)-linguistic variation. For further discussion of problems associated with interactions in VARBRUL, see Young and Yandell 1999 and Bayley 2001.

Another major challenge for sociolinguistics lies in finding appropriately sophisticated frameworks with which to understand the patterns that probabilistic analyses reveal—frameworks with adequate insight and explanatory power. The patterns revealed by our Oprah Winfrey study need explanation in many areas. Here we address just two: What are the *cognitive* patterns and processes through which such patterns of intra-speaker variation arise? And what are the *social* mechanisms and constructs that condition the observed behavior?

4.4.4.2 The Cognitive: What Is Variation? Our analysis of Winfrey's monophthongization patterns gives a good indication of their characteristics. Her orientation toward the person she is talking about (an important component of which is the person's ethnicity) affects the likelihood (and probably the degree) of monophthongization. Monophthongization is further influenced by the phonetic environment and by the lexical frequency of the word it appears in.

So what are the cognitive implications of these findings? What is Winfrey doing when she style-shifts? Models of speech production do not currently account for sociophonetic variation, even though this is a large part of what people do when they produce speech. One component of speech is clearly the continuous signaling of social identity and orientation. In order to satisfactorily begin to model this process, sociolinguists and those who work on speech will need to combine efforts.

One promising interpretation is that producing a phoneme (or word) involves the activation of a *distribution of phonetically detailed remembered examples* that characterize that phoneme (or word). More prototypical or central exemplars will be easiest to access, because of their central status in the distribution; and particularly frequent examples will also be easy to access, because of their high resting activation level. Exemplar theories of speech production and perception have been developed by, among others, Pierrehumbert (2001a, in press) for production and Johnson (1997b,c) for perception. Exemplar models are promising candidates for modeling sociophonetic effects because they do not treat variation as noise; on the contrary, variation is central and is inherently coded in lexical representations. Such models would appear to provide a

natural explanation for the involvement of lexical frequency in style shifting, as well as for why intraindividual style shifting echoes the interindividual social distribution of variables to which a speaker has been exposed (Bell 1984).

In Pierrehumbert's (2001a) implementation of exemplar theory, the selection of a phonetic target is modeled as random selection from a cloud of exemplars associated with the appropriate category. This models many social effects well, because "although social and stylistic factors may select for different parts of the exemplar cloud in different situations, the aggregate behavior of the system over all situations may be modeled as a repeated random sampling from the entire aggregate of exemplars" (Pierrehumbert 2001a, 145). Pierrehumbert demonstrates how a model with fully remembered exemplars can account for the fact that frequent words lead historical leniting changes and can model the timecourse of certain types of phonological merger.

The implementation is modified in Pierrehumbert, in press, so that production does not involve the specific selection of an exemplar, but rather can be heavily biased by activated exemplars. Exemplars are weighted and can be activated to different degrees in different contexts. Weighting can be affected by sociostylistic register and by contextual and attentional factors.

Goldinger (2000), Kirchner (in press), and Bybee (2001) also advocate exemplar-based models for speech production. And results reported by Goldinger (1997), Niedzielski (1999), Strand and Johnson (1996), and Whalen and Sheffert (1997), among others, provide strong evidence that social and speaker-specific information is not only stored, but also actively exploited in speech perception. Such results are highly consistent with models that include an exemplar-based level of representation, and they are very difficult to account for in models in which detailed exemplars are not stored.

Docherty and Foulkes (2000) have attempted to situate a discussion of sociophonetic variation in an exemplar model of lexical representation. Such a model accounts nicely for other patterns of variance such as coarticulation, connected speech processes, background noise effects, and intra- and interspeaker variability, and so, as Docherty and Foulkes point out, this seems a natural place to start. One of their central questions is "how phonology stands in the face of the variable aspects of a speaker's performance . . ." (p. 112). It would certainly seem that modeling sociophonetic variation would be a crucial test of the degree to which any

model of phonetic or phonological production and perception succeeds. However, it is not just models of speech that could benefit from such an understanding. Sociolinguists' understanding of the factors that are involved in style shifting, both linguistic and social, and the potential and possible ways in which they interact, would be deeply enriched by a clear understanding of the mechanisms through which this variation is represented and produced.

Resolving the nature of the cognitive status of probability distributions found in sociolinguistic studies would certainly make researchers' understanding and modeling of these phenomena more sophisticated and open new doors for analysis and explanation. By embedding studies of language variation in an understanding of language perception, production, and reproduction, researchers can start to consider how the observed probability distributions may come about, and how they might propagate, spread, and be manipulated in different social contexts for different social ends.

4.4.4.3 The Social: What Is Style? In the exemplar-theoretic view outlined above, social information that is interpretable by a listener is automatically stored with the exemplar, made more robust with repetition, and crucially linked to the actual instances of use of a particular variant. The proposal that linguistic categories, targets, and patterns are gradually built up through incremental experience with speech is entirely compatible with a view of the social world that relies on gradually built up social categories that emerge from the experiences that surround individuals as social actors. Just as there are no preset categories in phonology, and phonemes are abstracted from statistical patterning of the input (see Pierrehumbert, this volume, for extensive supporting evidence), so are social patterns abstracted and recovered from the same input.

We underscore the importance of interpretability by the listener. Within both the linguistic and the social world, young learners or foreign language speakers may not be equipped to fully understand the category composition of the stimuli to which they are exposed. It is only with repeated exposure that a child or a nonnative speaker can develop a robust enough model to incorporate and interpret new examples.

Because the development of an exemplar-based model proceeds example by example, it is important to look not only at overall distributions and gross statistical generalizations, but also at the micropatterning of individual instances. Understanding the flow of on-line discourse and its

relationship to robustness for both linguistic and social categories is an urgent task for sociolinguistics. Earlier, we mentioned that many of the social categories that researchers assume as given are not discrete, but may be treated as discrete for the purposes of statistical convenience. By supplementing statistical methods with qualitative analysis, we have exemplified one possible way to investigate how categories are built up in naturalistic contexts.

4.5 Conclusion

The use of probabilistic methods has led to important breakthroughs in sociolinguistics and has played an extremely important role in shaping the study of language variation and change. An important challenge for the future will be to move toward a more unified understanding of how subtle, gradient patterns of variation affect and are affected by cognitive, linguistic, and social structures, while always remembering that choices made for the analyst's convenience (such as treating monophthongization or ethnicity as binomial variables) are not pure mirrors of discrete categories in the world. We believe that the strongest theory of the interaction of language and society is a probabilistic theory, yet we encourage probabilistic sociolinguistic scholars to go beyond current methods: uncollapse what has been collapsed, and look for finer-grained social-theoretic explanations within what is uncovered in aggregate patterning.

Appendix

This appendix lists the different referees for the segments analyzed. Individuals were coded as "African-American" or "non-African-American." The "zero referee" cases involve segments in which the discourse is not focused on a specific individual.

Roseanne Barr, F actor	non-African-American
Halle Berry, F actor	African-American
George Clooney, M actor	non-African-American
Bill Cosby, M actor	African-American
Cindy Crawford, F model	non-African-American
Celine Dion, F musician	non-African-American
Mia Farrow, F actor	non-African-American
Daisy Fuentes, F actor	non-African-American
Kathy Lee Gifford, F actor	non-African-American
Delma Heyn, F writer	non-African-American
Lauren Hutton, F actor	non-African-American

Michael Jordan, M basketball player	African-American
Jackie Onassis, F celebrity	non-African-American
Brooke Shields, F actor	non-African-American
Will Smith, M actor/musician	African-American
Steven Spielberg, M movie director	non-African-American
Tina Turner, F musician	African-American
F Guest who dreams of having a house	African-American
"Gratitude Moments"	zero referee
"Oprah's Birthday"	zero referee
"How to Avoid Getting Scammed"	zero referee
"House to Style" (how to have more of it)	zero referee
"Oprah's Favorite Things"	zero referee

Note

The authors would like to acknowledge Rens Bod, Janet Pierrehumbert, and Kie Zuraw for extensive comments and suggestions. Malcah Yeager-Dror provided helpful guidance and Matt Loughren helped with references. All errors and omissions remain our own.