Variation in Linguistic Systems

Tying together work on a number of languages and linguistic varieties in different locales, this book provides students and researchers with a convenient, unified overview of variationist analysis in linguistics. *Variation in Linguistic Systems* takes a theoretical and quantitative approach to the study of variation in language, focusing on the role of language-internal constraints on variation and the relation of linguistic variation to linguistic theory. It introduces the basic concepts of variationist linguistics, and includes key discussions of different types of variation, multivariate analysis with GoldVarb, variation in sound and grammatical systems, language change and language contact.

Here is an ideal textbook for an introductory course on variation, as well as a useful resource for scholars with some background in linguistics who are interested in the study of language variation and its relation to the wider field of linguistics.

James Walker is Associate Professor of Linguistics in the Department of Languages, Literatures, and Linguistics at York University in Toronto.

Variation in Linguistic Systems

James A. Walker

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2010
by Routledge
270 Madison Avenue, New York, NY 10016

Simultaneously published in the UK
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Typeset in Sabon by
RefineCatch Limited, Bungay, Suffolk
Printed and bound in the United States of America on acid-free paper by
TO BE SUPPLIED

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
A catalog record has been requested for this book

ISBN10: 0–415–99067–X (hbk)
ISBN10: 0–415–99068–8 (pbk)
ISBN10: 0–203–85420–9 (ebk)

ISBN13: 978–0–415–99067–7 (hbk)
ISBN13: 978–0–415–99068–4 (pbk)
ISBN13: 978–0–203–85420–4 (ebk)

For Mom and Dad

Contents

Figures

Tables

Linguistic Corpora and Abbreviations

Unless otherwise indicated, all of the examples used in this book are cited verbatim from recordings of sociolinguistic interviews with actual speakers. Each example is identified by the corpus abbreviation, the speaker number, and the time index on the recording or the line number in the transcription of the interview. The following corpora are used, listed with their abbreviation and the reference that details the constitution of the corpus. I gratefully acknowledge permission to use examples from corpora on which I was not one of the principal investigators.

African Nova Scotian English	AN	Poplack and Tagliamonte (1991, 2001)
Bequia English	BQ	Meyerhoff, Sidnell and Walker (in preparation)
Ex-Slave Recordings	ES	Bailey, Maynor and Cukor-Avila (1991)
Montreal English	MQ	Poplack, Walker and Malcolmson (2006)
Nova Scotian English	NS	Poplack and Tagliamonte (1991, 2001)
Ottawa Hull French	OH	Poplack (1989)
Quebec City English	QC	Poplack, Walker and Malcolmson (2006)
Samaná English	SA	Poplack and Sankoff (1987)
Toronto English	TO	Hoffman and Walker (in press)

Acknowledgments

This book has a long gestation, going back to my graduate student days at the University of Ottawa in the mid-1990s, when I was frustrated at having to piece the principles of variationist linguistics together from reading packages and word of mouth. In graduate seminars and workshops I have taught at York University, the University of São Paulo and the University of Edinburgh over the past nine years, I have tried to develop a coherent "storyline" for teaching the analysis of linguistic variation, but I still relied largely on reading packages and word of mouth. I didn't take the idea of writing a book on this topic seriously until a dinner at the Human Bean in Edinburgh in 2006, when Miriam Meyerhoff and Jennifer Smith encouraged me to finally do it.

The cliché about standing on the shoulders of giants also happens to be true. I owe many thanks to my former academic supervisor at the University of Ottawa, Shana Poplack, not only for giving me thorough training in the principles and praxis of variationist analysis, but also for providing me with the example of a productive and critical scholar. Her influence is evident is every page of this book. I've also been privileged to be able to bend the ears of some of the foremost scholars in the field of variationist linguistics, including (of course) Bill Labov, as well as Bob Bayley, Henrietta Cedergren, Sandra Clarke, Greg Guy, Paul Kerswill, Ruth King, Salikoko Mufwene, David Sankoff, Sali Tagliamonte and Walt Wolfram.

This book has benefited enormously from critical reading and help from various people. Ronald Beline Mendes, Michol Hoffman and Panayiotis Pappas very kindly supplied me with examples from their data. Students in my undergraduate and graduate courses served as guinea pigs for the chapters on quantitative analysis and GoldVarb, and I thank them for their feedback. Paul Kerswill provided comments on the chapter on phonetic and phonological variation. Two anonymous reviewers provided substantial and thought-provoking comments that greatly improved the manuscript. The biggest thanks go to Miriam Meyerhoff and Rena Torres Cacoullos, who tirelessly read and commented on most of the book, as well as giving me encouragement when my confidence was flagging. Of course any remaining shortcomings in the book are my own responsibility.

Special thanks go to the people at Routledge, especially Louise Semlyn and Ivy Ip, who gently but insistently pushed me on. I'm extremely grateful to them for their understanding while I went through a difficult personal year that delayed completing the book. The support and encouragement of my family cannot be understated, especially from my parents Lois and Tony, who despite having more serious things to worry about kept asking me, "Is your book finished yet?"

1 Introduction

> Unfortunately, or luckily, no language is tyrannically consistent. All grammars leak.
>
> (Sapir 1921: 38)

1.0 Introduction

Variation is a pervasive fact of language. Every time we speak, we make choices that shape the language we use and that influence the linguistic choices of other speakers. Despite the fact of variation, it is often viewed as a problem in linguistics. Sapir's double-edged lament, cited above, is not uncommon in descriptive and theoretical studies, in which variation is acknowledged only in footnotes or passed over in silence. In linguistics programs, the discussion of variation is often relegated to the last few weeks of introductory courses or taught in upper-year elective courses. As students, many of us often wondered whether there is any connection between linguistics and linguistic variation.

Over the past forty years, William Labov and his students (or, in my case, a student of his student) have developed a quantitative research paradigm that seeks to incorporate variation into the scientific study of language. This research paradigm is often described as sociolinguistics, though this term is misleading in several ways. As Labov himself has pointed out (1972), the use of the term sociolinguistics (as opposed to simply linguistics) implies that there could be a science of language that does not take into account the social dimensions of linguistic behavior. In addition, there are other research paradigms described as sociolinguistics (the sociology of language, the ethnography of communication, discourse analysis, language policy and planning, and so on) that are not quantitative and/or that address rather different types of research questions (see, for example, Coulmas 1997). Finally, the use of the term sociolinguistics implies an exclusive focus on social considerations, such as sex/gender, social class, ethnicity, and so on. While such considerations obviously constitute a large part of the study of linguistic variation, a glance at the literature reveals what Guy (1993) calls a "Janus-like" concern with both the social and linguistic aspects of variation. For these reasons, I prefer to use the term variationist linguistics to refer to this research paradigm.

1.1 About the Book

The focus of this book is the linguistic side of the variationist method, which concerns the conditioning of linguistic variation by language-internal constraints and the relationship between linguistic variation and linguistic theory. While there are a number of good introductions to and overviews of sociolinguistics (Chambers 2008, Coulmas 1997, Trudgill 2000), they either devote little space to discussing linguistic variation or focus entirely on social factors. There are also a number of detailed "case studies" that deal with linguistic variation in specific research locales or address particular issues, such as the history of African American English (Poplack 2000, Poplack & Tagliamonte 2001, Wolfram & Thomas 2002). However, because of the relatively narrow focus of these studies, they may not be well known outside of their subfields or be of more general interest. Other books provide good guidebooks for fieldwork, analysis and/or statistics (Baayen 2008, Milroy & Gordon 2003, Paolillo 2002, Tagliamonte 2006a) but contain little or no discussion of linguistic factors. From another perspective, recent work has started to address the needed dialogue between linguistic theory and variationist linguistics (Adger 2006, Henry 1995), but this work tends to proceed from the direction of linguistic theory, sometimes in venues that are inaccessible to a wider audience, and may be daunting to those who are not well-versed in the latest theoretical developments.

This book is an attempt to fill a gap in the field, tying together work on a number of different languages and linguistic varieties in different locales to provide a unified discussion of the linguistic side of the study of linguistic variation. Although I have tried to include studies of a variety of languages, readers may note an unfortunate bias in favor of English. Part of this bias stems from the focus of my own work, which I have drawn upon heavily to provide illustrative examples of the ideas developed in this book. This bias also stems from the concentration of variationist studies on English and a handful of European-origin languages, such as (Canadian) French, (New World) Spanish and (Brazilian) Portuguese. There has been increasing interest in extending the variationist method to other languages, but these studies have yet to reach a critical mass to rival that of the other languages. However, I hope that this is something that will change over time.

The main objective of this book is to provide students and researchers with a convenient, unified overview of variationist linguistic analysis. It is intended to be suitable not only as the main textbook for an advanced undergraduate or introductory graduate course (perhaps supplemented by additional readings and original studies), but also as a general introduction for scholars with some background in linguistics who are interested in the study of linguistic variation and its relation to the wider field of linguistics. The book is not intended as a how-to manual or a statistics reference, though one chapter discusses the specifics of multivariate analysis with GoldVarb in some detail. This chapter is included in order to provide more background on interpreting the figures and tables presented in subsequent chapters and is also intended to stand on its own as a guide to using GoldVarb. While I have made an effort to cite the major works on linguistic variation, the book is not intended as a comprehensive review of the variationist literature. The principle guiding my selection of studies to serve as examples to illustrate theoretical or methodological points is ease of exposition. For this reason, I have drawn heavily on my own work and some better-known studies are passed over for detailed discussion in favor of lesser-known work. In citing works, I have tried to strike a balance between giving credit where credit is due and not overburdening the reader with excessive in-text citations. Interested readers should consult the works cited and the suggested readings provided at the end of each chapter for further details on specific studies.

1.2 Structure of the Book

The book is divided into nine chapters dealing with different methodological and theoretical aspects of variationist linguistics. Although the chapters build on each other, each includes an introduction recapping the content of the preceding chapter and outlining the structure of the chapter and a conclusion summarizing the main points of the chapter, to allow for the use of individual chapters depending on the reader's level of experience and background. In this chapter, we discuss the goals and structure of the book. Chapter 2 discusses different types of variation and introduces the basic concepts of variationist linguistics, including variables and variants, the principle of accountability, the importance of defining the variable context, and the difference between form-based and function-based approaches to studying linguistic variation. Chapter 3 discusses the analysis of linguistic variation in detail, contrasting categorical rules with variable rules, modeling relationships between variants, methods of calculating relative frequencies, testing hypotheses and methods for determining whether differences in frequency are statistically meaningful. Chapter 4 discusses the use of GoldVarb, a computer program that figures prominently in variationist linguistics, and proceeds step by step through multivariate analysis, including preparing token files and conducting single- and multi-factor analysis, as well as discussing the limitations of GoldVarb and methods for overcoming them. Chapter 5 provides a detailed discussion of variation in sound systems, including different types of phonetic and phonological variables, defining the variable context for these variables, methods for measuring phonetic variation and the different constraints conditioning the variation. Chapter 6 provides an overview of variation above the level of phonology, including different types of grammatical variable, approaches to the problem of defining the variable context and the different types of factor that condition grammatical variation. Chapter 7 applies the principles of variationist linguistics to issues in language change and grammaticalization, using linguistic conditioning to test different models of change. Chapter 8 focuses on the linguistic consequences of language contact, using the variationist method to answer questions of linguistic system membership in adult second language acquisition, convergence and pidgin/creole studies. Chapter 9 explores ways in which linguistic variation can be accommodated within linguistic theory. It is hoped that the book will spur discussion in all of these areas.

1.3 Further Reading

Chambers, J.K. 2008. *Sociolinguistic Theory*. Revised edition. Oxford and Malden, MA: Blackwell.

Chambers, J.K., Peter Trudgill and Natalie Schilling-Estes, eds. 2002. *The Handbook of Language Variation and Change*. Oxford and Malden, MA: Blackwell.

Paolillo, John C. 2002. *Analyzing Linguistic Variation*. Stanford, CA: CSLI Publications.

Tagliamonte, Sali. 2006. *Analysing Sociolinguistic Variation*. Oxford and Malden, MA: Blackwell.

2 Variation and Variables

2.0 Introduction

The previous chapter outlined the themes we will explore in this book. However, before we embark on an overview of studies of linguistic variation, we need to clarify what we mean by “variation”, since this term is used in linguistics in a number of different senses, not all of which are relevant for our purposes. In this chapter, we begin by discussing the different definitions of variation and defining the sense in which it is used in this book. We then introduce a number of basic concepts in the analysis of linguistic variation, which it will be important to understand before proceeding to the following chapters. We begin by defining what we mean by “variation”, before introducing the concept of the variable and variants. We discuss the principle of accountability and the importance of defining the variable context. We also consider the difference between form-based and function-based approaches to defining the variable context.

2.1 What is “Variation”?

In its broadest sense, variation refers to differences in linguistic form. In this sense, languages obviously differ from each other on a number of different levels.

Most salient and perhaps most trivial are lexical and phonological differences. Speakers are most conscious of linguistic differences in words and sounds. Listeners identify whether I am speaking English or Spanish by whether I refer to my household pet as [kʰæt] or [gɑto]. Some languages have sounds that other languages do not have. For example, English has an interdental fricative [θ], as in *think* and *bath*, which presents problems for English learners whose first language is French or Chinese, which do not have this sound. Conversely, English does not have a velar fricative [x], which presents a problem for English speakers who learn German, which does have this sound, as in *Bach*. Spanish distinguishes five vowels (/i/, /e/, /ɑ/, /o/, /u/), while Arabic distinguishes only among three (/i/, /ɑ/ and /u/).

Grammatical differences (that is, differences in morphology and syntax) are no less important, though they are apparently much less salient. Speakers rarely if ever comment on grammatical differences between languages, which take a number of forms. First, languages differ according to basic word order. For instance, as the example sentences in (2.1) show, English and Chinese are Subject–Verb–Object languages, Japanese and Tamil are Subject–Object–Verb languages, and Gaelic and Arabic are Verb-Subject-Object languages.

(2.1) Subject–Verb–Object
a. English the-woman$_{\text{Subject}}$ saw$_{\text{Verb}}$ the-children$_{\text{Object}}$
b. Chinese fùniü$_{\text{Subject}}$ kànjian$_{\text{Verb}}$ xiaohair$_{\text{Object}}$
Subject–Object–Verb
c. Japanese onna-ga$_{\text{Subject}}$ kodomo-o$_{\text{Object}}$ mita$_{\text{Verb}}$
d. Tamil[1] peɲ$_{\text{Subject}}$ kuɻandai-gaɭ-ai$_{\text{Object}}$ paar-tt-aaɭ$_{\text{Verb}}$
Verb–Subject–Object
e. Gaelic chunnaic$_{\text{Verb}}$ a’bhean$_{\text{Subject}}$ a’chlann$_{\text{Object}}$
f. Arabic[2] ra’aat$_{\text{Verb}}$ al-mar’aat$_{\text{Subject}}$ al-awlad$_{\text{Object}}$

Languages also differ according to the way that relations between the subject and verb are indicated morphologically. As shown in Table 2.1, the ending of the verb in Russian changes according to the subject in number and person, whereas the verb in Chinese does not. These differences, which I will refer to as cross-linguistic or interlinguistic variation, constitute the subject matter of linguistic typology. Cross-linguistic variation may occur across dialects (interdialectal) or even between individual speakers of the same language (idiolectal). When linguists speak of “variation”, this type of variation is usually what they are referring to.

In linguistics, we normally assume that interlinguistic variation holds across but not within languages, dialects or idiolects. For example, we do not expect speakers of English to vary between Subject-Verb and Verb-Subject word order. However, this type of variation does occur. Speakers of English sometimes use subject-verb order (2.2a) and verb-subject order (2.2b).

(2.2) a. We are going to the movies.
b. Are we going to the movies?

Table 2.1 Paradigm of present-tense subject-verb agreement for the verb *speak* in Russian and Chinese.

	Russian		**Chinese**	
	singular	plural	singular	plural
1st person	ja govorju	my govorim	wǒ shuō	wǒmen shuō
2nd person	ty govoriʃ	vy govorite	nǐ shuō	nǐmen shuō
3rd person	on(a) govorit	oni govorjat	tā shuō	tāmen shuō

Speakers of Brazilian Portuguese pronounce /o/ sometimes as [o] (e.g. *boca* [ˈboka] "mouth") and sometimes as [u] (e.g. *baço* [ˈbɑsu] "spleen"). In English, the plural morpheme /-s/ sometimes occurs as [s] (*cats* [kæts]), sometimes as [z] (*dogs* [dɑgz]), and sometimes as [əz] (*houses* [ˈhaʊzəz]).

Although the examples in the preceding paragraph represent variation ("differences in linguistic form"), we tend not to think of them as variation because we can provide linguistic explanations for this variation. The variation between Subject-Verb and Verb-Subject order in English (along with a change in intonation) corresponds to statements and questions. The variation between [o] and [u] in Brazilian Portuguese occurs in stressed and unstressed syllables, respectively. The variation in the English plural depends on the nature of the preceding consonant: [s] with preceding voiced consonants, [əz] with preceding sibilant consonants, and [z] everywhere else. In fact, the goal of linguistic analysis is to explain apparent variation, in one of two ways.

First, differences in form (phonetic, phonological, morphological, syntactic) may be explained by differences in meaning (lexical, grammatical, pragmatic). The change from Subject-Verb to Verb-Subject order in English signals a change from statement to question. In Chinese, changing the place of articulation of the initial consonant of a word changes the meaning of that word: if we change *p̲àng* to *t̲àng*, the meaning changes from "fat" to "hot". The difference in place of articulation is distinctive, or phonemic: [p] and [t] are phonemes of Chinese. In the Russian examples in Table 2.1, a change in the ending of the verb indicates a change in the person and number of the subject. In English, adding /-d/ to the end of [dræg] changes the reference of the verb to past tense. Thus, /-d/ in English is a morpheme indicating past tense. In conducting linguistic analysis, we normally assume that the relation between form and meaning is symmetrical, or one-to-one, a view expressed most succinctly by Dwight Bolinger (1977: x): "The natural condition of language is to preserve one form for one meaning and one meaning for one form." This assumption of form-meaning symmetry entails that any change in form is necessarily accompanied by a change in meaning.

Where changes in form cannot be explained by changes in meaning, we may try to correlate them with changes in the linguistic context. For example, we can state that "/o/ in Brazilian Portuguese is pronounced as [u] when it occurs in an unstressed syllable", which we might formalize as a rewrite rule, as in (2.3). This rule expresses the fact that the allophones of /o/, [o] and [u], are in complementary distribution: that is, there is no environment in which the two allophones can both occur.

(2.3) $$\text{o/} \rightarrow [\text{u}] \,/\, \frac{\qquad\quad}{[-\text{stress}]}$$

Similarly, we can state that "the English plural marker is pronounced [s] after voiceless consonants, [əz] after sibilant consonants and [z] elsewhere", which we could formalize as a distributional statement, as in (2.4), or as a set of rewrite rules, as in (2.5).[3]

(2.4)

$$\begin{tikzcd} & {[\text{s}]/\underset{[-\text{voice}]}{\text{C}} + \_\_\_\_} \\ /\text{z}/ \arrow[ur] \arrow[r] \arrow[dr] & {[\text{əz}]/\underset{[+\text{sibilant}]}{\text{C}} + \_\_\_\_} \\ & {[\text{z}] \text{ elsewhere}} \end{tikzcd}$$

(2.5) a. $/\text{z}/ \rightarrow [\text{əz}]/\underset{[+\text{sibilant}]}{\text{C}} + \_\_\_\_$

b. $\text{z}/ \rightarrow [\text{s}]/\underset{[-\text{voice}]}{\text{C}} + \_\_\_\_$

The crucial point about these statements, whether formalized as rewrite rules or as distributions, is that they make deterministic statements about differences in form. That is, every time there is a change in the linguistic context, there is a change in form. In conducting linguistic analysis, if the formulation of a rule does not predict the observed distribution of forms, we can reformulate the rule or distributional statement, or we can search for additional elements of the linguistic context until we can make a deterministic statement.

However, in many cases we may search in vain for elements of the linguistic context that allow us to make a deterministic statement. For example, English speakers sometimes pronounce words like *singing* as [ˈsɪŋɪŋ] or [ˈsɪŋɪn] and words like *west* as [wɛst] or [wɛs]. Spanish speakers sometimes pronounce words like *entonces* "then" as [enˈtonses] or [enˈtonseh]. In such cases, there is a difference in linguistic form but no change in the linguistic context and no apparent change in meaning. The most common response to this situation in linguistics is to label the forms as being in free variation, implying that the change in form is completely random and unpredictable (what we refer to as the null hypothesis of linguistic variation). Another response is to suggest that such variation is not conditioned by the linguistic context but rather by elements external to the linguistic system, such as the social context. In other words, while there is no change in linguistic meaning, there may be a change in social meaning.

This type of variation, which we will refer to simply as linguistic variation, is the subject of this book. From now on, when we refer to linguistic variation, we refer to changes in linguistic form without (apparent) changes in linguistic meaning for which we cannot make deterministic statements. In the following sections, we address the question of which changes in form count as linguistic variation (and which do not), and where such variation can (and cannot) occur.

2.2 Variables

Beginning in the 1960s, William Labov initiated a research program that aimed to study so-called "free" variation systematically, by correlating the variable realization of different phonological forms with social differences among speakers, stylistic differences among situations, and differences in linguistic context. His early work on English in Martha's Vineyard in Massachusetts (Labov 1963) and the Lower East Side in New York City (Labov 1966) has since been continued and expanded by his associates, students and students of students, in different languages and in various locales around the world (see Chambers, Trudgill and Schilling-Estes 2002 for a recent overview). This work has developed a coherent set of methodological principles for analyzing linguistic variation, which taken together we will refer to as variationist linguistics.

These studies have consistently demonstrated the futility of completely eliminating variation from the analysis of language. Approaches contemporaneous with Labov's original work tried to account for variation as a mixture of different lects, (internally invariant) linguistic systems, at the level of the community (Bailey 1973; Bickerton 1971; DeCamp 1971), or code-switching between different linguistic systems at the level of the individual. However, studies of variation have shown that linguistic variation exists even at the level of the individual speaker. Even controlling for social, stylistic and linguistic contexts, individual speakers still exhibit variable linguistic behavior. Rather than retaining the assumption of form-meaning symmetry, variationist linguistics recognizes that the relation between form and meaning may be asymmetrical: one meaning may be conveyed by several forms, and one form may correspond to different meanings. This form-function asymmetry is referred to by Labov (1972) as inherent variability: variation is an inherent property of human language, one that linguistic analysis should take into account rather than trying to eliminate.

Variationist linguistics depends crucially on the concept of the variable, which may be compared to the concept of the phoneme. As with the phoneme, the variable is an abstract construct, not something that we ever actually hear. Rather, what we hear are its overt manifestations: with phonemes, we hear allophones; with variables, we hear variants. There are also notational conventions associated with variables: just as we normally indicate phonemes using the notational convention of angled brackets /x/, we indicate variables with parentheses (x). For both allophones and variants, we generally use brackets: [x_1], [x_2], [x_3]. For example, the variation between forms like *singing* and *singin'* constitutes the English variable (ing), whose variants are the velar [ɪŋ] and alveolar [ɪn] realizations. The variation between the final sound in *entonces* and *entonceh* constitutes the Spanish variable (s)-aspiration, with the variants [s] and [h]. The variation between forms like *west* and *wes'* constitutes the English variable (t/d)-deletion, with an overt variant, [t] or [d], and a null variant. Throughout this book, we will see many more examples of variables at different levels of the linguistic system in different languages.

However, there are several important differences between the phoneme and the variable. A phoneme may consist of a single allophone, but a variable must consist of at least two variants (otherwise it would not be variable!). Another important difference has to do with distribution: while different allophones of a phoneme must never occur in the same linguistic context, variants of a variable may (indeed, must) occur in the same context. However, as we will see in Chapter 3, although variants do not occur in complementary distribution, each may occur with greater or lesser frequency than other variants of the same variable when certain elements of the linguistic or social context are present.

Let us (provisionally) define a variable as "different ways of saying the same thing", with the "different ways" being the variants (we consider the question of "the same thing" in §2.3). Any time the speaker has a choice between forms, we can begin to investigate whether this choice constitutes a variable.

Speakers have choices between linguistic forms at a number of different levels of language. We have already seen examples from phonetics and phonology: (ing) and (t/d)-deletion in English and (s)-aspiration in Spanish (we explore these variables in more detail in Chapter 5). The choice of lexical item may constitute another type of variable: whether English speakers say *couch* or *sofa* (or, for older Canadians, *chesterfield*), they refer to the same piece of furniture. Subject-verb agreement is another place where speakers may have a choice: whether English speakers say (2.6a) or (2.6b), they are referring to the same thing.

(2.6) a. There were foxes around there. (QC9: 362)
b. But there was a lot of foxes. (QC9: 361)

In many varieties of English (such as African American English and nonstandard northern British English), third person plural occurs sometimes as standard *they go* and sometimes as nonstandard *they goes*. In Brazilian Portuguese, many speakers who use the new first person plural pronoun *a gente* (< "the people") vary between the more standard third person singular ending (2.7a) and the nonstandard first person plural ending (2.7b).

(2.7) a. *a gente vai* "we (lit. the people) go (3rd sg.)"
b. *a gente vamos* "we go (1st pl.)"

In French, speakers have the choice between expressing the future as a verb ending (2.8a) or an auxiliary (2.8b).

(2.8) a. Si on parle seulement que français, ça prendr-a pas de temps. (OH10: 3389)
if one speaks only that French, that take-FUTURE not of time
"If we only speak French, that will take no time."
b. Peut-être ça va prendre mille ans, mais . . . (OH4: 2842)
can-be that goes to-take thousand years, but
"Maybe it's going to take a thousand years, but . . ."
(Poplack & Turpin 1999)

Speakers often have a choice between realizing or deleting elements of the sentence. In Spanish, the subject pronoun may be overt or null (2.9).

(2.9) Y entonces pero ella era tan gritona, que cuando ella lo decía, Ø lo decía tan y tan fuerte.
"And then but she was so loud, that when she would say it, (she) would say it so loud."
(Cameron 1993: 314–15)

The English complementizer *that* is sometimes omitted when introducing a subordinate clause, as shown in (2.10).

(2.10) a. Everyone thinks Ø I'm from Montreal. (MQ67: 1778)
b. Anybody that comes here knows that I don't speak it. (QC57: 1408)

Speakers of English may also choose between different strategies for reporting speech: a verb of saying (2.11a), the verb *go* (2.11b) or a newer form involving *be like* (2.11c).

(2.11) a. I said, "I'm not the weather man." (TO14: 40)
b. I go, "How many times have I covered for you?" (TO14: 732)
c. He was like, "Oh if you're not, I'm going to." (TO14: 486)

Variables such as those illustrated in (2.6) to (2.11) are discussed in Chapter 6.

All of these examples of variation involve different forms, at different levels of the linguistic system—phonetic/phonological, lexical, morphological, syntactic and discursive/pragmatic—but are they really "different ways of saying the same thing"?

2.3 The Variable Context

If a variable reflects choices among forms that speakers make, we have to determine exactly where these choices are possible (and where they are not). Here we come to a central methodological problem in the analysis of linguistic variation, that of defining the variable context (also known as the envelope of variation): that is, which forms count as variants of a variable? Defining the variable context is an important step in the analysis of linguistic variation—perhaps the most important step—because it affects all subsequent analysis, it affects the results obtained and, ultimately, it affects the interpretation of those results. Part of defining the variable context involves determining which forms alternate with each other. We need to take into account not only the form we are interested in, but also the other forms with which that form varies. This consideration, known as the principle of accountability, means that before we begin calculating rates of occurrence, we need to know how to calculate those rates, which depends on how the variable context is defined. A number of approaches have been taken to defining the variable context, but all of them can be broadly classified as what I will call form-based and function-based.

Form-based approaches to defining the variable context begin by noting that two (or more) forms that are (roughly) equivalent in meaning alternate with each other. This is relatively uncontroversial at the level of phonology. Whether you say *singing* or *singin'*, it refers to the same activity. Since this variation occurs in all forms of unstressed final *-ing* (i.e. Verb-*ing, nothing, everything*, but not *ring*), we can define the variable context of (ing) as "word-final unstressed *-ing*". Similarly, the variable context of (s)-aspiration can be defined "word-final /s/", and that of (t/d)-deletion as "word-final /t/ or /d/ in a consonant cluster". We may also want to define the variable context for a phonological variable on the basis of a particular lexical item or class of lexical items containing a particular phoneme. For example, many studies of regional dialectology have noted alternate pronunciation of *schedule* with an initial [sk] or [ʃ] (the former usually associated with North America and the latter with the United Kingdom). Studies of vowel variables typically make reference to classes of phoneme, such as English /ɛ/ (including *pet*, *left*, *ten*, and so on) or "front lax vowels" (we will discuss this in more detail in Chapter 5). In fact, form-based approaches are common in studies of phonological and lexical variables, typically defined in terms of a particular structural configuration. However, it is possible to take a form-based approach to grammatical variables. For example, Montreal French alternates between the auxiliary verbs *être* "to be" and *avoir* "to have" in forming the past tense, as shown in (2.12) (Sankoff & Thibault 1977).

(2.12) a. J'ai rentré à cinq heures, j'ai été opérée le lendemain matin à dix heures et demie.
I-have entered at five hours, I-have been operated the next-day morning at ten hours and half
"I went in at five o'clock, I was operated on the next morning at 10:30."
b. Je suis rentrée juste la veille de l'opération à cinq heures, j'ai été opérée le lendemain matin à dix.
I am entered just the night-before of the-operation at five hours, I-have been operated the next-day morning at ten
"I went in the night before the operation at five o'clock, I was operated on the next morning at ten."
(Germaine C., 1984; Sankoff & Thibault 1977)

Similarly, other studies have noted alternation in English in expression of the future, with *will* or *going to* as robust alternatives, as shown in (2.13).

(2.13) a. They've almost been married twenty-five years, I think next year will be their twenty-fifth. (TO27: 554)
b. You know, there's gonna be one to Ottawa in November, I think. (TO27: 905)

The problem with form-based approaches to grammatical variation is that it becomes more difficult to meet the requirement of semantic equivalence. Given the possibility of form-meaning symmetry, it needs to be demonstrated that grammatical alternatives are in fact "different ways of saying the same thing". We discuss this question in more detail in Chapter 6.

Rather than enumerating a set of equivalent forms, we may proceed in the opposite direction, defining a particular linguistic function and noting all the different forms that convey that function. This function-based approach is more common for grammatical variables than for phonological or lexical variables, in part because of the problems inherent in defining the variable context for grammatical variation. For example, we could define the function of "reference to future time" and note all of the forms that convey this function: *will* (2.13a), *going to* (2.13b), simple present (2.14a), present progressive (2.14b), and others.

(2.14) a. And then we go to Stratford or something like that. (TO27: 906)
b. Actually, they're coming up to the cottage this weekend. (TO27: 509)

At the same time, we would exclude from the variable context those occurrences of "future" forms that do not refer to future time, such as *will* in habitual contexts (2.15a) and the progressive in situations of ongoing activity (2.15b).[4]

(2.15) a. He'll golf on a- a- a Friday and I'll get things ready and
then he'll pack up the car and we'll go off and, you know,
maybe stay a week. (TO27: 972)
b. We're getting to the point we don't need that.
(TO27: 873)

At the level of discourse, we could define the function of (QUOTATIVE) and note all of the different forms used for this function: *say*, *go*, *be like*, etc., as in (2.11). The function-based approach does not solve the problem of equivalence noted above, but it does sidestep the issue by reformulating the notion of semantic equivalence. We will discuss these issues in further detail in Chapter 6.

In addition to determining where the variation can occur, defining the variable context involves determining where the variation cannot occur. Regardless of whether we take a form-based or function-based approach, we want to exclude contexts in which we do not find variation. For example, in defining the variable context for (ing) we find a number of word-final occurrences of *-ing* in which the alveolar variant never occurs: that is, words like *ring* and *bring* never occur as *rin'* and *brin'*. These categorical (i.e. 0 percent or 100 percent) contexts can be excluded by reformulating the variable context as "word-final unstressed *-ing*". We also want to exclude particular lexical items that, while not categorical, may be highly associated with one of the variants. For example, in terms of (t/d)-deletion, the word *and* almost never occurs with the final [d]. Including such forms would inflate the overall rate of deletion beyond the average. Similarly, we do not want to include contexts in which we cannot reliably determine which of the variants occurred. In collecting tokens of (t/d)-deletion, if someone says *half past ten* [hæfpæastɛn], it is unclear whether the final [t] in *past* is pronounced, since it coalesces with the onset of the following syllable.

Once we have defined the variable context, we can then proceed to extract occurrences, or tokens, of the variable from the data. Studies may take as their data recordings (or transcriptions of recordings) of natural speech or historical or online texts. The most common source of data in variationist research is the sociolinguistic interview, in which informants are encouraged to speak for an hour or two on topics of interest to them (see Labov 1984 and Tagliamonte 2006a for more details). The variable context guides us as to which forms may be included and which should not be included. Once a sufficient number of tokens have been extracted, we can then begin to classify the tokens as to the social and linguistic context. This is the subject of Chapter 3.

2.4 Conclusion

In this chapter we have established the groundwork for the analysis of linguistic variation. We began by defining linguistic variation as “differences in linguistic form without (apparent) change of meaning”, distinguishing this use of the term from other uses common in linguistics. We introduced a number of analytical concepts that we will refer to throughout the book. A fundamental concept is the variable and its variants, a formal or functional unit and its overt realizations. We noted that variation can occur at various levels of the linguistic system. Two important considerations in the analysis of linguistic variation are defining the variable context and obeying the principle of accountability, according to which we must determine where speakers have a choice between forms and what forms count as variants of the same variable. We outlined two broad approaches to defining the variable context: form-based and function-based. The definition of the variable context guides the extraction of data for analysis. In the next chapter, we will discuss the analysis of linguistic variation in more detail.

2.5 Further Reading

Guy, Gregory R. 1993. The quantitative analysis of linguistic variation. In D. Preston (ed.), *American Dialect Research*. Amsterdam: Benjamins, 223–49.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Wolfram, Walt. 1993. Identifying and interpreting variables. In D. Preston (ed.), *American Dialect Research*. Amsterdam: Benjamins, 193–221.

3 The Analysis of Linguistic Variation

3.0 Introduction

The previous chapter laid the groundwork for the analysis of linguistic variation. We defined linguistic variation as “differences in linguistic form without (apparent) changes in meaning”. We also introduced the analytic construct of the variable, which we defined as “different ways (variants) of saying the same thing (the variable context)”. We saw that variables can occur at a number of different linguistic levels: phonetics or phonology, morphology, syntax, the lexicon and discourse. We introduced the principle of accountability, which requires that we examine not only the variant of interest to us but also its relative frequency with respect to all of the other variants of the same variable. In this respect, the definition of the variable context, the place where the speaker has a choice between forms, assumes a central position in the analysis of linguistic variation. How we define the variable context determines which forms we include in the analysis, how we calculate relative frequencies and, ultimately, how we interpret the variation.

Although one of our goals is to calculate the overall relative frequency of each variant, more important is determining the contribution made by elements of the (linguistic) context to the choice of each variant. The goal is not to search for categorical distributions of each variant (all vs. nothing), but rather to look for a change in its relative frequency across different contexts (more vs. less). Thus, the analysis of linguistic variation is inherently quantitative and requires recourse to standard statistical procedures, which allow us to determine whether the differences in relative frequencies are meaningful, as well as the relative contribution of each contextual element to the occurrence of a particular variant.

In this chapter, we discuss the analysis of linguistic variation in detail. We begin by returning to the formulation of linguistic processes as rules, which leads us to contrast categorical rules with variable rules. We consider the relations that may exist between variants of the same variable and how to decide among competing scenarios. We outline the methods involved in calculating relative frequencies of variants, not only in the simple cases but also when there are more complex relations between variants. We discuss the considerations involved in testing hypotheses and determining whether differences in frequency can be considered meaningful.

3.1 Variable Rules

In Chapter 2, we noted that the goal of linguistic analysis is to explain variation, either through differences of meaning or through different distributions of forms according to context. The expected outcome is a deterministic statement about the distribution of forms, such that any difference in meaning or context necessarily correlates with a difference in form. The formulation of a deterministic statement constitutes a categorical rule, since it applies 100 percent of the time. As we have seen, however, there are many cases in which we cannot make a deterministic statement about the distribution of linguistic forms. In these situations, the forms are said to be in free variation, the formulation of which constitutes an optional rule, since it applies randomly.

In early work on African American English, Labov (1969) pointed out that the application of so-called "optional" rules is rarely if ever completely random. If we took a sufficiently large number of observations of the application (and non-application) of an optional rule, we could make a quantitative generalization about the likelihood that the rule applies. Rather than calling such a rule optional, he proposed that we redefine it as a variable rule (indicated with angled brackets < >). A variable rule is similar to a categorical rule in containing a structural description to which it applies, such as the (categorical) English plural rule discussed in Chapter 2 (reproduced here as (3.1)).

(3.1) a. $$/\text{z}/ \rightarrow [\text{əz}] \,/\, \frac{\text{C}}{[\text{+sibilant}]} + \underline{\quad\quad}$$

b. $$/\text{z}/ \rightarrow [\text{s}] \,/\, \frac{\text{C}}{[-\text{voice}]} + \underline{\quad\quad}$$

However, in contrast to a categorical rule, which occurs 100 percent of the time, a variable rule includes as part of its formulation a particular rate of application. For example, we could formulate the variables of English (t/d)-deletion and Spanish (s)-aspiration as in (3.2) and (3.3), respectively.

(3.2) $$\text{(t/d)-deletion:} \begin{bmatrix} \text{C} \\ -\text{cont} \\ -\text{son} \\ +\text{cor} \end{bmatrix} \rightarrow \langle \text{Ø} \rangle \,/\, \text{C}\ \underline{\quad}\ \#$$

(3.3) $$\text{(s)-aspiration:}\ [\text{s}] \rightarrow \langle \text{h} \rangle \,/\, \underline{\quad}\ \#$$

The rule of (t/d)-deletion would be paraphrased as "variably delete [t] and [d] in a consonant cluster at the end of a word", while the rule of (s)-aspiration would be paraphrased as "variably change [s] to [h] at the end of a word". (We will discuss how to determine the rule's rate of application when we discuss calculating relative frequencies in §3.3.)

At the time that Labov developed the notion of the variable rule, the dominant linguistic paradigm was Transformational-Generative Grammar (Chomsky 1965; Chomsky & Halle 1968), which viewed linguistic processes as a set of ordered rules that derive (spoken) Surface Structure from an underlying Deep Structure. Linguistic theory at that time was concerned with formulating rules and determining the orderings of rules to account for different languages. As a result, early work on variable rules was also concerned with the details of rule formulation and rule ordering.[1] A legacy of this history is the use of the phrase variable rule analysis to refer to the analysis of linguistic variation. In practice, most variationist studies nowadays tend not to concern themselves with the specific formulation of rules. In fact, as we will see, in many cases it is not necessary or even appropriate to formulate a variable as a rule. Whether or not we view the variation as resulting from a rule in the traditional transformational-generative sense, the variable rule remains one useful way (among several) of thinking about the choices that speakers make, regardless of the linguistic mechanism(s) we assume to underlie those choices.

3.2 Modeling Relations Between Variants

If we take the view that variation is not completely random but results from speaker choices, we must consider the linguistic processes giving rise to the variation. In other words, what kinds of relationship exist between variants of the same variable?

Let us start with the simplest case, a variable (v) with two variants, [v_1] and [v_2]. We have already seen examples of this case, such as English (t/d)-deletion and Spanish (s)-aspiration. There are two possible relationships: [v_1] is the input (the underlying form) and [v_2] is the output (the surface form), as in (3.4a); or [v_2] is the input and [v_1] is the output, as in (3.4b).

(3.4)		**Variable Rule**	**Input**		**Output**
	a.	VR-1	/v_1/	→	<v_2>
	b.	VR-1'	/v_2/	→	<v_1>

The type of rules represented by VR-1 and VR-1' depend on what linguistic process we assume. If [v_1] represents an overt variant and [v_2] a null variant, VR-1 is a deletion rule, as in (3.5a), and VR-1' is an insertion rule, as in (3.5b).

(3.5)	a.	(t/d)-deletion	/t/d/	→	<Ø> / C __ #
	b.	(t/d)-insertion	/Ø/	→	<t/d> / C __ #

How do we decide which formulation of the rule is appropriate? Which variant is the input and which is the output? The variable rule notation does not help us to answer this question. Rather, we must use linguistic and statistical reasoning to decide. In the case of (t/d), it makes more sense to assume that an underlying consonant is variably deleted word-finally than that a word-final segment is variably inserted. For example, [t] and [d] are not variably inserted at the end of words for which there is no evidence of a final consonant otherwise: that is, *sand* varies with *san'*, but *man* never varies with *mand*. (We will discuss statistical procedures for making this decision in later chapters.)

Let us move on to a more complicated case, a variable with more than two variants. As an illustration, we use an issue first identified by Henrietta Cedergren in her 1972 study of Panama City. As we noted in Chapter 2, a final /s/ in a word like *entonces* may be pronounced as [s] or [h]. There is also a third variant, in which the final segment is deleted altogether, [Ø]. Thus, there seem to be two rules at work, a rule of aspiration and a rule of deletion. One possibility, formulated in (3.6), is that these two rules operate independently: that is, aspiration and deletion both apply to word-final /s/ without regard to each other:

(3.6) a. (s)-aspiration /s/ → <h> / ___ #
b. (s)-deletion /s/ → <Ø> / ___ #

An alternative view is that these rules are not completely independent, but interact with each other in some way. We could view aspiration and deletion as different degrees of consonantal weakening, representing a continuum: *s* → *h* → Ø. We could reformulate the deletion rule such that it takes as its input not the underlying [s] but rather the output [h] of the (s)-aspiration rule, as formulated in (3.7).

(3.7) a. (s)-aspiration /s/ → <h> / ___ #
b. (h)-deletion /h/ → <Ø> / ___ #

In this scenario, the aspiration rule "feeds" the deletion rule. How do we decide which of these two scenarios is appropriate? As we saw earlier, when we needed to decide which of two variants represents the input to the variable rule, this decision needs to be based on what makes the most sense linguistically. In this case we can appeal to other processes in the history of Spanish. In the development of Spanish from Latin, the fricative [f] weakened to [h], which was later deleted. For example, Latin *filius* "son" became Spanish *hijo*, and although modern Spanish retains the *h* in spelling, it is not pronounced. Thus, we may view a rule that weakens the fricative [s] to [h], which feeds a rule that deletes [h], as a (variable) synchronic reflection of a related historical process. (We will show later that quantitative reasoning can also help us to resolve this question.)

Note that we can only make such arguments on a language-by-language basis. For example, while Brazilian Portuguese also has a variable rule of (s)-deletion (e.g. Naro 1981), it has no rule of (s)-aspiration. Therefore, the deletion rule of Brazilian Portuguese would be formulated as in (3.8), different from that of Spanish.

(3.8) (s)-deletion/s/ → <Ø> / __ #

Spanish aspiration and deletion provide a clear-cut case of evidence for rule ordering, but is such a model appropriate for other variables? For example, as we saw in Chapter 2, reference to future time in English is variably expressed by different grammatical constructions: *will, going to*, the simple present and the present progressive, as illustrated in (3.9).

(3.9) a. They've almost been married twenty-five years, I think next year will be their twenty-fifth. (TO27: 554)
b. You know, there's gonna be one to Ottawa in November, I think. (TO27: 905)
c. And then we go to Stratford or something like that. (TO27: 906)
d. Actually, they're coming up to the cottage this weekend. (TO27: 509)

We could posit a variable (FUTURE) with four variants (we will return to the question of defining a variable context for grammatical variables in Chapter 6). Is there an argument for thinking of these variants as the result of (multiple) variable rules, with one variant as the input, and ordering of the rules? This view does not make much sense linguistically: it would be hard to argue that one of these forms is more "basic" to the future than the others, and a rule that derives one form from the other would be rather unnatural. What this variable is really saying, then, is that, when referring to future time, speakers have a choice between different grammatical options.

Although the English future illustrates the problem with conceptualizing all choices as variable rules, it nevertheless also involves the same issues involved in multiple-variant variables. Instead of conceptualizing the relationship between variants as a set of (ordered) rules, we could argue that there are two variable contexts, one nested inside the other. These are shown as decision trees in Figures 3.1 and 3.2. In Figure 3.1, Spanish speakers have two choices: they can aspirate /s/ or not; if they aspirate /s/, they can delete the resulting /h/ or not. As shown in Figure 3.2, English speakers have two sets of choices for expressing the future: they can choose the present or not; if they choose the present, they have a choice between simple and progressive; if they do not choose the present, they have a choice between *will* and *going to*. Again, the decision of

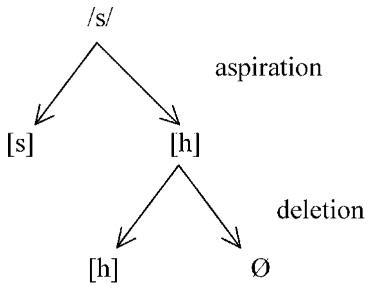


Figure 3.1 Decision tree for variants of Spanish (s)-aspiration.

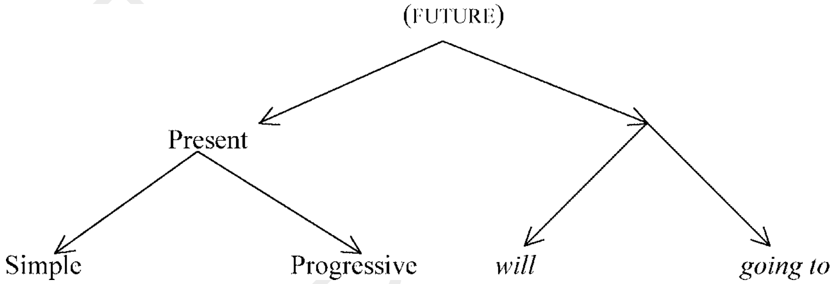


Figure 3.2 Decision tree for variants of future temporal reference in English.

whether this is the correct way of conceptualizing the relation between variants needs to be decided on linguistic and statistical grounds.

It may seem that we are spending an inordinate amount of time considering the relationships between variants. However, keep in mind that variationist analysis is an exercise in linguistic analysis. As we will see, the type of relationship we assume has consequences for the calculation of relative frequencies.

3.3 Calculating Frequencies

In Chapter 2, we noted that the purpose of defining the variable context is to specify which forms vary with each other. This step in the analysis is of crucial importance because it determines which forms we include in the analysis (variants) and how we calculate frequencies for each of the variants. By comparing frequencies across (social or linguistic) contexts, we can determine whether these contexts have an effect on the frequency with which the rule applies.

Let us again begin with the simplest situation, a variable with two variants. The frequency of whichever variant we choose will always be relative to the frequency of the other variant. For example, to calculate the frequency of the alveolar variant of (ing), the principle of accountability states that we need to know not only the number of times that the alveolar variant occurs, but also the number of times that it could occur but does not (i.e. how many times the velar variant occurs):

$$(3.10)\quad \frac{\#\text{occurrences}}{\#\text{occurrences} + \#\text{non} - \text{occurrences}} = \frac{\#\ \textit{-in'}}{\#\ \textit{-in'} + \#\ \textit{-ing}}$$

(Note that the denominator is equivalent to the variable context.) Assume that we observe 50 occurrences of (ing) in a conversation and we find that 20 occur as *-in'* and 30 as *-ing*. Using the formula above, we arrive at the following distribution:

$$(3.11)\quad \frac{\#\text{occurrences}}{\#\text{occurrences} + \#\text{non} - \text{occurrences}} = \frac{20}{20 + 30} = \frac{20}{50} = 40\%$$

Thus, the relative frequency of the alveolar variant in our data is 40 percent. To arrive at the relative frequency of the velar variant, we subtract the relative frequency of the alveolar variant from 100 percent: 100% – 40% = 60%.

What if there are more than two variants? The way we calculate relative frequencies will depend on our assumptions about the relationship between the variants (see §3.2). Returning to the example of Spanish (s)-aspiration and (s)-deletion, assume that we observe 50 occurrences of word-final /s/ and we find 20 occurrences as [s], 20 as [h] and 10 as [Ø]. In the first scenario, we assume that aspiration (3.12a) and deletion (3.12b) operate on the underlying [s] independently of each other.

(3.12) a. (s)-aspiration:

$$\frac{\#\text{occurrences}}{\#\text{occurrences} + \#\text{non} - \text{occurrences}} =$$

$$\frac{\#\ [\text{h}]}{\#\ [\text{h}] + (\#\ [\text{s}] + \#\ [0])} = \frac{20}{20 + (20 + 10)} = \frac{20}{50} = 40\%$$

b. (s)-deletion:

$$\frac{\#\text{occurrences}}{\#\text{occurrences} + \#\text{non} - \text{occurrences}} =$$

$$\frac{\#\ [0]}{\#\ [0] + (\#\ [\text{s}] + \#\ [\text{h}])} = \frac{10}{10 + (20 + 20)} = \frac{10}{50} = 20\%$$

Thus, we can state that the rate of (s)-aspiration is 40 percent and the rate of (s)-deletion is 20 percent. (The other 40 percent is occurrences of [s]: that is, forms that have not undergone aspiration or deletion.)

If we assume a feeding relationship between aspiration and deletion, we need to calculate the rates slightly differently. In this scenario, since all tokens of [Ø] are [h] at some point in the derivation, we must assume that they are first aspirated before being deleted: [s] → ([h] → [Ø]). (Note that we can think about this in terms of rules or as two nested variable contexts, as in Figure 3.1.) This translates into a different set of calculations for aspiration (3.13a) and deletion (3.13b).

(3.13) a. (s)-aspiration:

$$\frac{\#\text{occurrences}}{\#\text{occurrences} + \#\text{non} - \text{occurrences}} =$$

$$\frac{(\#\,[\text{h}] + \#\,[0])}{(\#\,[\text{h}] + \#\,[0]) + \#\,[\text{s}]} = \frac{(20 + 10)}{(20 + 10) + 20} = \frac{30}{50} = 60\%$$

b. (h)-deletion:

$$\frac{\#\text{occurrences}}{\#\text{occurrences} + \#\text{non} - \text{occurrences}} =$$

$$\frac{\#\,[0]}{\#\,[0] + \#\,[\text{h}]} = \frac{10}{10 + 20} = \frac{10}{30} = 33\%$$

Note that the relative frequencies of aspiration and deletion are different in each scenario. If we do not assume that the rules are ordered with respect to each other (or that the variable contexts are nested), the frequency of aspiration is 40 percent and the frequency of deletion is 20 percent. If we assume rule ordering (or nesting of variable contexts), the frequency of aspiration is 60 percent and the frequency of deletion is 33 percent. From this comparison, we can see that the assumptions we make about the relationships between variants not only imply different types of linguistic processes, but they also give rise to different quantitative results.

3.4 Testing Hypotheses

In analyzing linguistic variation, we not only want to calculate relative frequencies for each of the variants, but we also want to know whether particular contextual elements influence the choice of variant. If the variation is truly "free", contextual elements should have no influence over the choice of form: the frequency of each variant will remain (roughly) the same regardless of the linguistic context. This prediction, which constitutes the null hypothesis (H_0) of linguistic variation, in essence says that the variation is completely random (see Chapter 2). In order to disprove this hypothesis, we need to demonstrate that the linguistic context has an effect on the choice of variant. In contrast with other types of linguistic analysis, however, our goal is not to make a deterministic statement based on categorical distributions of the variants in each context, but rather to make probabilistic statements based on relative distributions of variants across contexts. That is, given the presence of a particular linguistic context, we predict a difference (increase or decrease) in the relative frequencies of variants.

The predictions that we make about the effect of contextual elements on the choice of variant represent hypotheses that may be drawn from a number of sources. Previous studies of the same variable may have discovered an effect. For example, since studies have found lower rates of (t/d)-deletion if the [t] or [d] occurs in a past tense form (e.g. *missed*) than if it is part of the preceding morpheme (e.g. *mist*) (e.g. Guy 1980), this is a hypothesis we could test. We may make predictions on the basis of a particular theory of language. For example, a functional theory of language predicts that meaningful elements are more likely to be retained. Thus, we may predict that forms of [s] in Spanish that serve to indicate the plural (e.g. *casas* "houses") are less likely to undergo deletion than forms in which the [s] is part of the preceding morpheme (e.g. *entonces* "then") (e.g. Poplack 1980a). We may already have an informal impression that frequencies are different in a particular context, which we wish to test quantitatively. The advantage of the variationist method is its "pretheoretical" nature (Laks 1992), in that it does not impose a particular theory on the analysis. The decision of which hypotheses to test depends on the model of language adopted by the researcher. In this sense, variationist linguistics is not inherently structuralist, formalist, generativist or functionalist: rather, we can use the variationist method to test structuralist, formalist, generativist or functionalist hypotheses. The only restriction is that these hypotheses must provide some sort of (socially or linguistically) meaningful explanation about the variation and, most importantly, should lend themselves to empirical investigation.

Although the statistical methods used to test hypotheses in the analysis of linguistic variation are no different from those used in psychology and the social sciences, there are some "in-house" terminological differences that may cause some confusion, so we should clarify them before proceeding (see also Sankoff 1988b). What we have been referring to as the variable is normally referred to in statistics as the dependent variable, the object whose behavior we are studying. Testing hypotheses involves operationalizing them as what are normally referred to in statistics as independent variables (consisting of values or levels), the contextual elements whose effect on the dependent variable we are testing. In variationist terminology, independent variables are factor groups, which consist of factors corresponding to different options within the factor group. For example, we may hypothesize that (t/d)-deletion is more likely to occur if the next sound is a consonant rather than a vowel (e.g. Guy

1980). We could then define a factor group called "following phonological context", consisting of the factors "consonant" and "vowel". In defining factor groups, it is important that the factors within each group be mutually exclusive. For example, if we made a finer division of the following phonological context as "consonant", "sonorant" and "vowel" (3.14a–c), a following [l] (3.14d) would present a problem, since it is both a consonant and a sonorant. We could solve this problem by redefining the "consonant" factor to "obstruent".

(3.14) a. We were the mixed kids. (TO.#: 19: 43)
b. The rest of us just turn' nineteen. (TO.F: 22: 15)
c. A nice home-cooked Italian meal. (TO.6: 54: 26)
d. We los' like two lifeguards. (TO.M: 22: 45)

Factors within each factor group must also be exhaustive: we must be able to code each token into one of the factors. For example, as currently defined, "following phonological context" would have difficulty if there was no sound following the variable (for example, at the end of an utterance, as in (3.15)). We could fix this problem by adding another factor to the factor group, "pause".

(3.15) . . . a ghost that floats around. (TO.7: 33: 50)

In the analysis of linguistic variation, we normally have more than one hypothesis about which contextual elements condition the variation. For example, we may hypothesize that (t/d)-deletion is affected not only by the following sound but also by the preceding sound (e.g. Guy & Boberg 1997), as well as whether or not the [t] or [d] serves to mark past tense (e.g. Guy 1980). This entails defining a number of factor groups for each variable. For (t/d)-deletion, we could define three factor groups to test our hypotheses:

(3.16)

Factor group	**Factors**	**Example**
Following phonological context	Consonant	*kept cool*
	Vowel	*kept out*
	Pause	*kept*
Preceding phonological context	Stop	*kept*
	Fricative	*left*
	Sonorant	*sent*
	Sibilant	*mist*
	Other consonant	*cold*
Morphological status	Past	*missed*
	Non-past	*mist*

Once we have operationalized our hypotheses as factor groups, we can begin extracting occurrences of the variable, or tokens, from the data (recorded conversations or sociolinguistic interviews, transcriptions of conversations, radio or TV broadcasts, historical documents, letters, and so on). Each token we extract needs to be classified or coded according to one of the factors for each of the factor groups that we have defined. When coding tokens, it is not as important whether a particular factor group represents the "right" hypothesis as it is to make consistent decisions. For example, if we coded a preceding [r] sometimes as "sonorant" and sometimes as "other consonant", the relative frequencies for the preceding phonological context would clearly be questionable.

3.5 Examining Distributions

Once a sufficient number of tokens have been collected and coded, we can examine the distribution of variants across factors by calculating the relative frequency of each variant within each factor and comparing frequencies across factors within the same factor group. To illustrate, we will use tokens of (t/d)-deletion taken from recorded conversations with speakers of English in Toronto (Hoffman & Walker, in press).

Table 3.1 shows the distribution of deleted and non-deleted tokens for each factor within the factor group. For each factor, the relative frequency of deletion is shown. Relative frequencies are calculated by dividing the number of deleted tokens by the total number of tokens in each factor. Table 3.1 seems to confirm our original hypothesis: deletion occurs at a much higher rate with a following consonant (66 percent) than with a following pause (37 percent) or vowel (32 percent).

A common mistake is to calculate the proportion of tokens across the factors instead of within each factor. For example, we could calculate the proportion of deleted tokens among consonants (671/1082 = 62%), vowels (270/1082 = 25%) and pauses (141/1082 = 13%). This would tell us the distribution of one variant (deletion) across different contexts, but it would tell us nothing about speaker choices, which can only be determined by calculating the proportion of variants within each factor.

Table 3.1 Distribution of variants of (t/d)-deletion by following phonological context in Toronto English.

Following Context	# Deleted	# Non-deleted	Total #	Rate of Deletion
Consonant	671	342	1013	66%
Vowel	270	588	858	32%
Pause	141	239	380	37%
Total #	1082	1169	2251	48%

3.6 Testing for Statistical Significance

The difference in relative frequency between following consonant and following pause or following vowel seems large, but is it meaningful? This difference could reflect random "noise" in our data, due to some other factor (or factors) that we have not considered. Answering this question requires recourse to tests of statistical significance.

The term significance is used in casual speech to refer to what we perceive as important or salient, but it has a specific definition in statistics. When we use statistical methods, we need to recognize that repeated measurements of the dependent variable will vary around an average value, and we may not be able to completely account for all of this variance. Nevertheless, we can figure out how likely it is that this variance is affected by the independent variable(s) we are investigating. The alternative, that the variance is not affected by the independent variable(s), is known as the null hypothesis (H_0). To test these competing hypotheses, we need to know how the data would be distributed if the null hypothesis were true: that is, if the distribution is truly random (the expected values). We then compare the expected distribution with the actual distribution (the observed values). The distance between the expected and observed values is usually expressed as a numerical value of variance (or deviation). If the variance is large enough (that is, if the observed and expected values are sufficiently different), we can conclude that the distribution is significant, and that the variation is likely affected by the independent variable.

To make this decision, we need one more piece of information. We need to know the minimum number of dimensions or parameters of the statistical model (such as the number of measurements or factors). This number, known as the degrees of freedom, corresponds to the number of pieces of information we need to know in order to know what all the values in our model are. A good analogy is a jigsaw puzzle: for a 500-piece jigsaw puzzle, you only need to fit together 499 pieces before you know where all 500 fit, since the position of the final piece can be predicted once the other 499 pieces are in place. So the degrees of freedom for a 500-piece jigsaw puzzle is 500 – 1 = 499.

Once we know the variance and the degrees of freedom, we can compare these values against known values for the null hypothesis. Based on this comparison, we can figure out how probable it is that the null hypothesis is true. This probability is expressed as a numerical value p. The lower the p-value, the less likely it is that the observed distribution is due to chance. Nowadays most statistical software packages will provide an exact value for p, but traditionally in statistics the p-value is expressed in relation to a cutoff point, such as 10 percent, 5 percent, or 1 percent. Below this cutoff point ($p < 10\%$, $p < 5\%$, $p < 1\%$ . . ., also expressed as $p < .10$, $p < .05$, $p < .01$, . . .), the variance is considered statistically significant. The cutoff point depends on how sure you want to be that the results are not due to chance. In medicine, the cutoff should obviously be set very low (1 percent or even less!). In linguistics, the cutoff tends to be set at 5 percent, though *p*-values slightly above that may be considered marginally significant.

There are a number of tests of statistical significance, but as an example we will use the chi-square test to determine whether the pattern in Table 3.1 (repeated in the top third of Table 3.2) is significant. The chi-square statistic is a numerical measurement of how much variance there is, or in other words, how much distance there is between the observed distribution and the expected distribution (the null hypothesis). First, we figure out the expected values for each cell in the table, using the formula shown in the middle third of Table 3.2 and the observed values. We then measure the difference between the expected and observed values by calculating a chi-square statistic for each cell, using the formula shown in the bottom third of Table 3.2. Adding up the chi-square values of all the cells to derive a total chi-square value, as shown at the bottom of Table 3.2, gives

Table 3.2 Chi-square test for Toronto (t/d)-deletion data.

Observed Values

	# Deleted	# Non-deleted	Row Total
Consonant	671	342	1013
Vowel	270	588	858
Pause	141	239	380
Column Total	1082	1169	2251

Expected Values

$$\frac{\text{(row total)} \times \text{(column total)}}{\text{(grand total)}}$$

	# Deleted	# Non-deleted	Row Total
Consonant	487	526	1013
Vowel	412	446	858
Pause	183	197	380
Column Total	1082	1169	2251

Chi-square Values

$$\frac{\text{(observed} - \text{expected)}^2}{\text{(expected)}}$$

	# Deleted	# Non-deleted	
Consonant	69.5	64.4	
Vowel	48.9	45.2	
Pause	9.6	9.0	
		Total Chi-square Value:	246.6

Degrees of freedom (*df*) = (number of columns – 1) × (number of rows – 1)
= (2–1) × (3–1)
= 2

Table 3.3 Percentage points of chi-square distribution (adapted from Woods, Fletcher & Hughes 1986: 301).

			p =			
Degrees of freedom:	**50%**	**10%**	**5%**	**1%**	**0.1%**	
1	.45	2.71	3.84	6.64	10.8	
~~2~~	~~1.39~~	~~4.61~~	~~5.99~~	~~9.21~~	~~13.8~~	► 246.6
3	2.37	6.25	7.82	11.3	16.3	
4	3.36	7.78	9.49	13.3	18.5	
5	4.35	9.24	11.1	15.1	20.5	

us a total chi-square value (variance) of 246.6. The degrees of freedom for the table (df = 2) are calculated using the formula underneath Table 3.2. As shown in Table 3.3 (adapted from Woods, Fletcher & Hughes 1986), using the degrees of freedom to look up the chi-square value in a chi-square table (available as an appendix in most statistics manuals or online resources) shows that the total chi-square value of 246.6 is well below the cutoff of p = 5% (actually, well below even p = 1%), which means that there is a less than 5 percent probability that the null hypothesis is true. We can therefore conclude that the following phonological context exerts a statistically significant effect (p < .05) on the occurrence of (t/d)-deletion.

Tests of statistical significance (such as the chi-square test) are useful and appropriate for testing the significance of each factor group individually. However, since there are normally multiple hypotheses about the influence of contextual factors on the variation, we need to consider the possibility that different factor groups may affect the variation simultaneously. Moreover, since factor groups may act together in various ways, we need to make use of statistical techniques that can account for the individual effects of each factor group when all of them are considered together.

3.7 Conclusion

This chapter provided a detailed discussion of the analysis of linguistic variation. We began by returning to the formulation of linguistic processes as rules, which led us to contrast categorical with optional rules. We refined the notion of optionality as the variable rule, a useful way of modeling speaker choices. We considered the types of relationship that exist between variants of the same variable and what this relationship says about the linguistic processes involved and whether there is evidence for ordering of variable rules. We outlined methods for calculating relative frequencies for two-variant and multiple-variant variables. We discussed the steps involved in testing hypotheses about the contextual factors conditioning the variation, and how we can determine whether such effects are statistically significant. We noted the need for statistical tests that allow us to consider the effects of multiple factor groups simultaneously. In the next chapter, we look at such tests.

3.8 Further Reading

Baayen, Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Bayley, Robert. 2002. The quantitative paradigm. In J.K. Chambers, P. Trudgill & N. Schilling-Estes (eds.), *The Handbook of Language Variation and Change*. Oxford: Blackwell, 117–41.

Paolillo, John C. 2002. *Analyzing Linguistic Variation: Statistical Models and Methods*. Stanford: CSLI Publications.

Tagliamonte, Sali. 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Young, Richard & Robert Bayley. 1996. VARBRUL analysis for second language acquisition research. In D. Preston (ed.), *Second Language Acquisition and Linguistic Variation*. Amsterdam/Philadelphia: Benjamins, 253–306.

4 Multivariate Analysis with GoldVarb

4.0 Introduction

In the previous chapter, we discussed the basic principles of the analysis of linguistic variation. We introduced the notion of the variable rule, a useful way of modeling speaker choices, and we discussed methods for calculating relative frequencies, using factor groups to test hypotheses about the effects of the linguistic context on the variation. We also provided a method for determining whether such effects are statistically significant, though we noted the need for further statistical tests that would allow us to consider the effects of several factor groups simultaneously.

In this chapter, we discuss multivariate analysis, using the program GoldVarb. GoldVarb is widely used in variationist analysis over other statistical programs (such as SPSS) because of availability free of charge for both PC and Macintosh computers and its relative user-friendliness. Most importantly, GoldVarb was developed specifically for the analysis of linguistic variation, in which data are often not distributed evenly across all factors and factor groups. Unevenly distributed data present problems for other types of multivariate analysis that are commonly used in psychology and sociology, such as ANOVA.

GoldVarb X (Sankoff, Tagliamonte & Smith 2005) is the most current version of the VARBRUL family of computer programs (starting with Cedergren & Sankoff 1974). Although GoldVarb is more user-friendly than other statistical programs, it is still necessary to learn the idiosyncrasies of the program in order to conduct quantitative analysis (both single-factor and multi-factor), as well as to understand its output, which commonly figures in variationist studies. In this chapter, we will proceed step by step through multivariate analysis in GoldVarb, though we will not discuss every detail of the operation of the program (readers are referred to the program's documentation). We begin by discussing how data are formatted for GoldVarb token files, before proceeding to a discussion of how to generate percentages using condition files. We then discuss multivariate analysis in GoldVarb, making use of the step-up/

step-down procedure. We will also discuss some of GoldVarb's limitations and techniques for working around them.

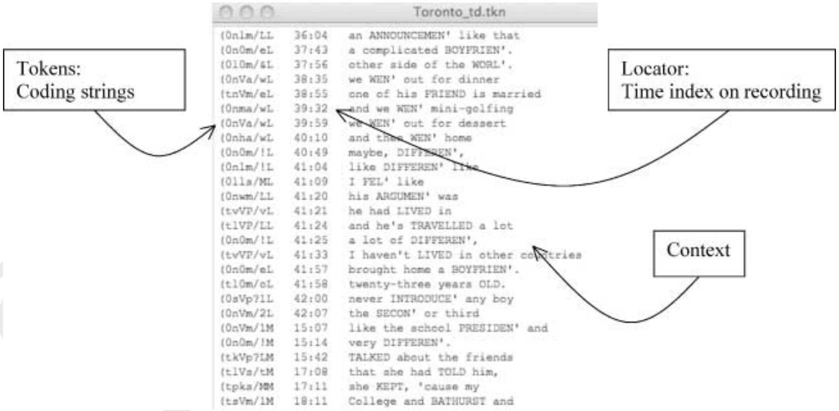
4.1 Token Files

In GoldVarb, data are stored in a token file (*.tkn). A token file is just a flat-text ASCII file, but GoldVarb requires that the data be formatted in a particular way. The program treats each line that begins with a left parenthesis as the beginning of a token and reads the following string of characters until the specified number of characters has been reached. It then looks for the next left parenthesis at the beginning of a line, which it treats as the next token. Each token consists of a string of codes (single characters, each of which represents a factor in the factor group), normally with the (dependent) variable as the first character in the coding string. A single-character code must be assigned to each factor before coding begins. The codes in (4.1) (using the factor groups developed in (3.14)) give an example of a possible set of coding instructions for (t/d)-deletion.[1]

(4.1)	**Factor group**	**Factors**	**Code**
	Variant	[t] or [d]	t
		Ø	0
	Following phonological context	Consonant	c
		Vowel	v
		Pause	p
	Preceding phonological context	Stop	t
		Fricative	f
		Sonorant	n
		Sibilant	s
		Other consonant	c
	Morphological status	Past	p
		Non-past	n

Although GoldVarb reads only the coding string and ignores everything else until it encounters the next left parenthesis at the beginning of a line, many researchers include a locator for each token (such as the time index or line number in the text) and the context of each token. Including a locator and the context is good practice in case you want to check your coding later. Figure 4.1 shows a fragment of the token file for our (t/d)-deletion data.[2] Each token begins with a left parenthesis and a coding string of seven characters, the first character being the (dependent) variable of (t/d)-deletion and the following six characters the codes for each factor group. Each token also includes a locator (here, the time index) and the context (the words immediately before and after the word containing the token).

There are different options for coding the data. The easiest option is to enter the data directly into the token file in GoldVarb. The main advantages

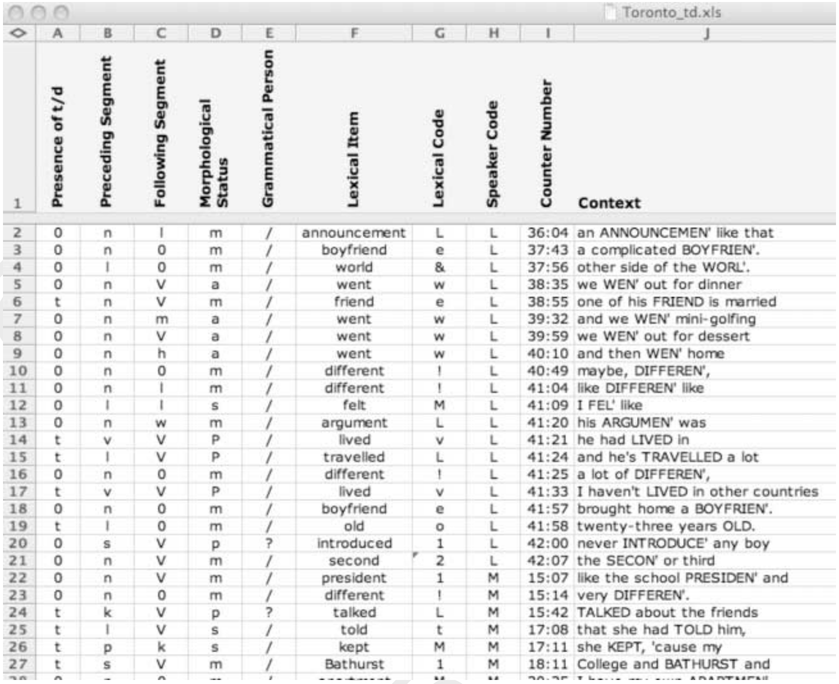


```
(0nlm/LL   36:04   an ANNOUNCEMEN' like that
(0n0m/eL   37:43   a complicated BOYFRIEN'.
(010m/&L   37:56   other side of the WORL'.
(0nVa/wL   38:35   we WEN' out for dinner
(tnVm/eL   38:55   one of his FRIEND is married
(0nma/wL   39:32   and we WEN' mini-golfing
(0nVa/wL   39:59   we WEN' out for dessert
(0nha/wL   40:10   and then WEN' home
(0n0m/!L   40:49   maybe, DIFFEREN',
(0nlm/!L   41:04   like DIFFEREN' like
(011s/ML   41:09   I FEL' like
(0nwm/LL   41:20   his ARGUMEN' was
(tvVP/vL   41:21   he had LIVED in
(t1VP/LL   41:24   and he's TRAVELLED a lot
(0n0m/!L   41:25   a lot of DIFFEREN',
(tvVP/vL   41:33   I haven't LIVED in other countries
(0n0m/eL   41:57   brought home a BOYFRIEN'.
(t10m/oL   41:58   twenty-three years OLD.
(0sVp?1L   42:00   never INTRODUCE' any boy
(0nVm/2L   42:07   the SECON' or third
(0nVm/1M   15:07   like the school PRESIDEN' and
(0n0m/!M   15:14   very DIFFEREN'.
(tkVp7LM   15:42   TALKED about the friends
(t1Vs/tM   17:08   that she had TOLD him,
(tpks/MM   17:11   she KEPT, 'cause my
(tsVm/1M   18:11   College and BATHURST and
```

Figure 4.1 Fragment of a token file for (t/d)-deletion.

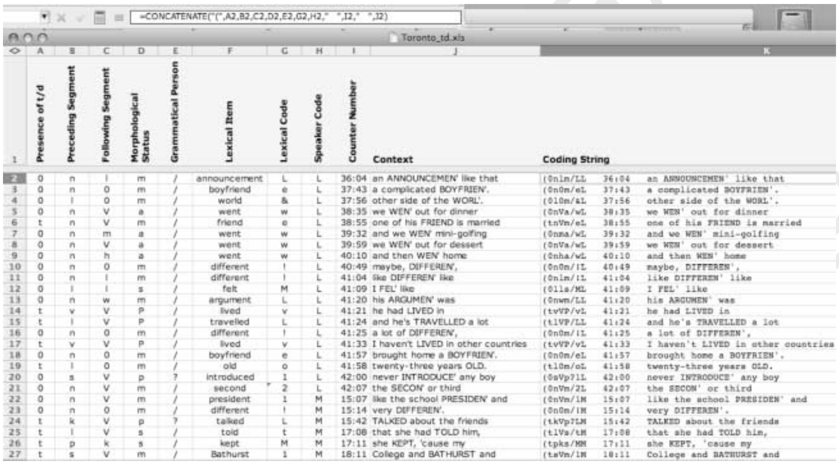
of this option are that no further formatting of the data is required, and at any time you can do a quick analysis of the data you have already coded. However, because the data in the token file are stored in a flat-text file, the main disadvantage is that you cannot sort the data while coding. This is unproblematic with factor groups containing few factors, but may present problems if the total number of factors is unknown (see the next paragraph). This method also requires you to code every factor group for each token as you are entering the data, which tends to be more time-consuming than coding by factor group, and may result in more coding errors.

Alternatively, you can code the data in another program and import it into GoldVarb. In a spreadsheet program such as Microsoft Excel, each row corresponds to one token and each column to one factor group, as shown in Figure 4.2, which contains the same fragment of the tokens as Figure 4.1. This method has the advantage of allowing you to sort the data while coding it, and it also avoids the need to come up with exhaustive codes for factor groups with a large number of factors before coding begins. For example, it is common to introduce a factor group for individual lexical items to test for lexical effects. Since there are potentially thousands of factors in this factor group and only 256 ASCII characters available as codes, we cannot code this factor group exhaustively in advance. However, as you can see from Figure 4.2, a spreadsheet makes it possible to enter individual lexical items in full (Column F) and assign the most frequent lexical items an individual code and the less frequent items a code of frequency (L = low, M = medium) (Column G). The disadvantage of this approach is that the tokens need to be modified to a format that GoldVarb can read. In Excel, you can use the CONCATENATE function to merge cells into a string which can then be imported into a GoldVarb token file. For example, in the Excel coding file shown in Figure 4.2, updated as Figure 4.3, we introduce a new column (Column



	A	B	C	D	E	F	G	H	I	J
1	Presence of t/d	Preceding Segment	Following Segment	Morphological Status	Grammatical Person	Lexical Item	Lexical Code	Speaker Code	Counter Number	Context
2	0	n	l	m	/	announcement	L	L	36:04	an ANNOUNCEMEN' like that
3	0	n	0	m	/	boyfriend	e	L	37:43	a complicated BOYFRIEN'.
4	0	l	0	m	/	world	&	L	37:56	other side of the WORL'.
5	0	n	V	a	/	went	w	L	38:35	we WEN' out for dinner
6	t	n	V	m	/	friend	e	L	38:55	one of his FRIEND is married
7	0	n	m	a	/	went	w	L	39:32	and we WEN' mini-golfing
8	0	n	V	a	/	went	w	L	39:59	we WEN' out for dessert
9	0	n	h	a	/	went	w	L	40:10	and then WEN' home
10	0	n	0	m	/	different	!	L	40:49	maybe, DIFFEREN',
11	0	n	l	m	/	different	!	L	41:04	like DIFFEREN' like
12	0	l	l	s	/	felt	M	L	41:09	I FEL' like
13	0	n	w	m	/	argument	L	L	41:20	his ARGUMEN' was
14	t	v	V	P	/	lived	v	L	41:21	he had LIVED in
15	t	l	V	P	/	travelled	L	L	41:24	and he's TRAVELLED a lot
16	0	n	0	m	/	different	!	L	41:25	a lot of DIFFEREN',
17	t	v	V	P	/	lived	v	L	41:33	I haven't LIVED in other countries
18	0	n	0	m	/	boyfriend	e	L	41:57	brought home a BOYFRIEN'.
19	t	l	0	m	/	old	o	L	41:58	twenty-three years OLD.
20	0	s	V	p	?	introduced	1	L	42:00	never INTRODUCE' any boy
21	0	n	V	m	/	second	2	L	42:07	the SECON' or third
22	0	n	V	m	/	president	1	M	15:07	like the school PRESIDEN' and
23	0	n	0	m	/	different	!	M	15:14	very DIFFEREN'.
24	t	k	V	p	?	talked	L	M	15:42	TALKED about the friends
25	t	l	V	s	/	told	t	M	17:08	that she had TOLD him,
26	t	p	k	s	/	kept	M	M	17:11	she KEPT, 'cause my
27	t	s	V	m	/	Bathurst	1	M	18:11	College and BATHURST and

Figure 4.2 Fragment of an Excel coding file for (t/d)-deletion.



=CONCATENATE("(",A2,B2,C2,D2,E2,G2,H2," ",I2," ",J2)

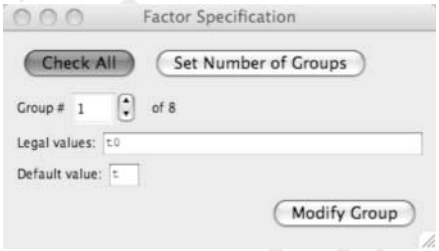
	A	B	C	D	E	F	G	H	I	J	K
1	Presence of t/d	Preceding Segment	Following Segment	Morphological Status	Grammatical Person	Lexical Item	Lexical Code	Speaker Code	Counter Number	Context	Coding String
2	0	n	l	m	/	announcement	L	L	36:04	an ANNOUNCEMEN' like that	(0nlm/LL 36:04 an ANNOUNCEMEN' like that
3	0	n	0	m	/	boyfriend	e	L	37:43	a complicated BOYFRIEN'.	(0n0m/eL 37:43 a complicated BOYFRIEN'.
4	0	l	0	m	/	world	&	L	37:56	other side of the WORL'.	(0l0m/&L 37:56 other side of the WORL'.
5	0	n	V	a	/	went	w	L	38:35	we WEN' out for dinner	(0nVa/wL 38:35 we WEN' out for dinner
6	t	n	V	m	/	friend	e	L	38:55	one of his FRIEND is married	(tnVm/eL 38:55 one of his FRIEND is married
7	0	n	m	a	/	went	w	L	39:32	and we WEN' mini-golfing	(0nma/wL 39:32 and we WEN' mini-golfing
8	0	n	V	a	/	went	w	L	39:59	we WEN' out for dessert	(0nVa/wL 39:59 we WEN' out for dessert
9	0	n	h	a	/	went	w	L	40:10	and then WEN' home	(0nha/wL 40:10 and then WEN' home
10	0	n	0	m	/	different	!	L	40:49	maybe, DIFFEREN',	(0n0m/!L 40:49 maybe, DIFFEREN',
11	0	n	l	m	/	different	!	L	41:04	like DIFFEREN' like	(0nlm/!L 41:04 like DIFFEREN' like
12	0	l	l	s	/	felt	M	L	41:09	I FEL' like	(0lls/ML 41:09 I FEL' like
13	0	n	w	m	/	argument	L	L	41:20	his ARGUMEN' was	(0nwm/LL 41:20 his ARGUMEN' was
14	t	v	V	P	/	lived	v	L	41:21	he had LIVED in	(tvVP/vL 41:21 he had LIVED in
15	t	l	V	P	/	travelled	L	L	41:24	and he's TRAVELLED a lot	(tlVP/LL 41:24 and he's TRAVELLED a lot
16	0	n	0	m	/	different	!	L	41:25	a lot of DIFFEREN',	(0n0m/!L 41:25 a lot of DIFFEREN',
17	t	v	V	P	/	lived	v	L	41:33	I haven't LIVED in other countries	(tvVP/vL 41:33 I haven't LIVED in other countries
18	0	n	0	m	/	boyfriend	e	L	41:57	brought home a BOYFRIEN'.	(0n0m/eL 41:57 brought home a BOYFRIEN'.
19	t	l	0	m	/	old	o	L	41:58	twenty-three years OLD.	(tl0m/oL 41:58 twenty-three years OLD.
20	0	s	V	p	?	introduced	1	L	42:00	never INTRODUCE' any boy	(0sVp?1L 42:00 never INTRODUCE' any boy
21	0	n	V	m	/	second	2	L	42:07	the SECON' or third	(0nVm/2L 42:07 the SECON' or third
22	0	n	V	m	/	president	1	M	15:07	like the school PRESIDEN' and	(0nVm/1M 15:07 like the school PRESIDEN' and
23	0	n	0	m	/	different	!	M	15:14	very DIFFEREN'.	(0n0m/!M 15:14 very DIFFEREN'.
24	t	k	V	p	?	talked	L	M	15:42	TALKED about the friends	(tkVp?LM 15:42 TALKED about the friends
25	t	l	V	s	/	told	t	M	17:08	that she had TOLD him,	(tlVs/tM 17:08 that she had TOLD him,
26	t	p	k	s	/	kept	M	M	17:11	she KEPT, 'cause my	(tpks/MM 17:11 she KEPT, 'cause my
27	t	s	V	m	/	Bathurst	1	M	18:11	College and BATHURST and	(tsVm/1M 18:11 College and BATHURST and

Figure 4.3 Fragment of an Excel coding file for (t/d)-deletion, with an added concatenation column.

K) and in the first row enter the formula shown in (4.2a), which generates the coding string shown in (4.2b). Copying the contents of this cell to the other cells in the new column generates a coding string for every token. This column can then be copied and pasted into GoldVarb.

```
(4.2)  a. =CONCATENATE("(",A2,B2,C2,D2,E2,G2,H2,"
          ",I2,"",J2)
       b. (0nlm/LL  36:04  an ANNOUNCEMEN' like that
```

One final step remains in preparing the token file, which is to instruct GoldVarb as to how many factor groups there are in the coding string, and what the legal values are for factors within each factor group. To do this, you use the Factor Specification window, as shown in Figure 4.4.[3]



Group # 1 of 8

Legal values: t0

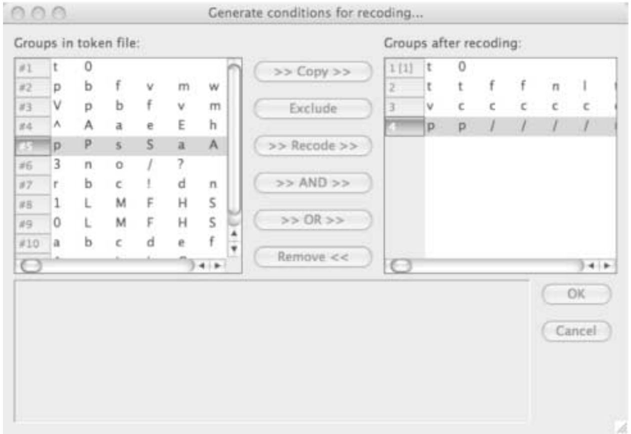
Default value: t

Figure 4.4 Factor specification window.

4.2 Calculating Frequencies

Once we have a token file, we need to tell GoldVarb how to analyze the data. To do this, we make use of a condition file (*.cnd), which contains a list of instructions to the program about which factor groups to include and how to analyze them. Most of the time we do not want to analyze tokens exactly the way they were coded. Rather, we want to reconfigure them as needed as we refine our hypotheses. We may want to combine factors within factor groups (making fewer distinctions), combine factor groups (turn two factor groups into one), exclude factors from a factor group, or not include factor groups in the analysis. Recoding the token file manually every time we wanted to change the configuration of the data would be very time-consuming (and might introduce coding errors!). Using condition files allows us to make changes to the way the data are analyzed without having to recode the token file. Ideally, once the token file has been coded, we do not want to make changes to it.

Like token files, condition files are flat-text files, but unlike token files, they can be generated by GoldVarb using a graphic user interface. Selecting "Recode Setup" in the "Tokens" menu opens the graphic interface, as shown on the left-hand side of Figure 4.5.[4] The factor groups as originally coded are listed on the left side of the window and the recoded factor groups appear on the right side. The function buttons in the middle of the



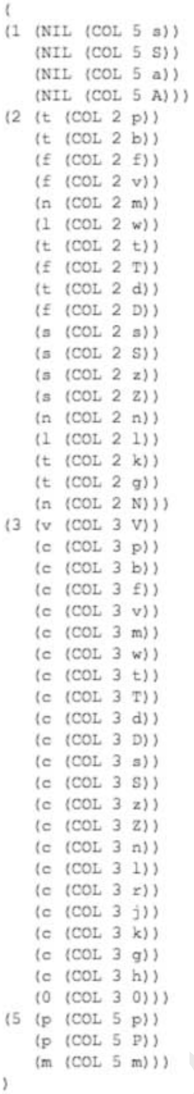
Generate conditions for recoding...

Groups in token file:

#1	t	0				
#2	p	b	f	v	m	w
#3	V	p	b	f	v	m
#4	^	A	a	e	E	h
#5	p	P	s	S	a	A
#6	3	n	o	/	?	
#7	r	b	c	!	d	n
#8	1	L	M	F	H	S
#9	0	L	M	F	H	S
#10	a	b	c	d	e	f

Groups after recoding:

1 [1]	t	0				
2	t	t	f	f	n	l
3	v	c	c	c	c	c
4	p	p	/	/	/	/



```
(
(1 (NIL (COL 5 s))
   (NIL (COL 5 S))
   (NIL (COL 5 a))
   (NIL (COL 5 A)))
(2 (t (COL 2 p))
   (t (COL 2 b))
   (f (COL 2 f))
   (f (COL 2 v))
   (n (COL 2 m))
   (l (COL 2 w))
   (t (COL 2 t))
   (f (COL 2 T))
   (t (COL 2 d))
   (f (COL 2 D))
   (s (COL 2 s))
   (s (COL 2 S))
   (s (COL 2 z))
   (s (COL 2 Z))
   (n (COL 2 n))
   (l (COL 2 l))
   (t (COL 2 k))
   (t (COL 2 g))
   (n (COL 2 N)))
(3 (v (COL 3 V))
   (c (COL 3 p))
   (c (COL 3 b))
   (c (COL 3 f))
   (c (COL 3 v))
   (c (COL 3 m))
   (c (COL 3 w))
   (c (COL 3 t))
   (c (COL 3 T))
   (c (COL 3 d))
   (c (COL 3 D))
   (c (COL 3 s))
   (c (COL 3 S))
   (c (COL 3 z))
   (c (COL 3 Z))
   (c (COL 3 n))
   (c (COL 3 l))
   (c (COL 3 r))
   (c (COL 3 j))
   (c (COL 3 k))
   (c (COL 3 g))
   (c (COL 3 h))
   (0 (COL 3 0)))
(5 (p (COL 5 p))
   (p (COL 5 P))
   (m (COL 5 m)))
)
```

Figure 4.5 Graphic interface for generating condition files in GoldVarb, with the resulting condition file.

window perform the operations. Factor groups can be included as is (COPY), changed (RECODE) or combined with other factor groups (AND or OR). Any factor groups not copied to the right side of the window are not included in the analysis. Factors may be excluded from factor groups (using EXCLUDE). Figure 4.5 shows a recoding of factor groups. Once the recoding is complete, click on OK.

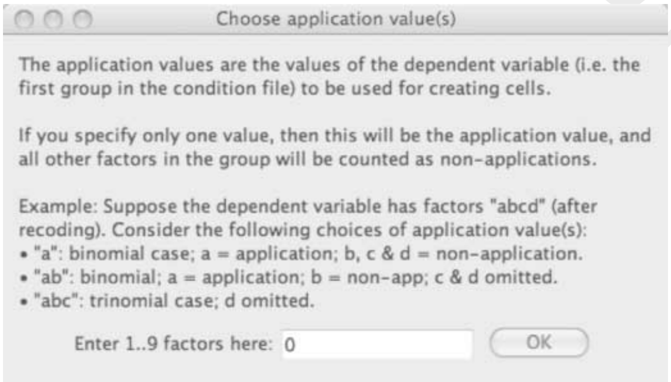
GoldVarb generates a condition file, an example of which is shown on the right-hand side of Figure 4.5. Instructions to the program can also be written directly into the condition file, without the graphic interface. However, this method requires learning the LISP programming language that the program uses to read the instructions. Although LISP is not complicated, starting with the more user-friendly graphic interface is recommended.

Once the token file and the condition file are ready, load the data into the computer's memory using "Load Cells to Memory" in the "Cells" menu. GoldVarb presents a window that asks us to specify an application value, as in Figure 4.6. The application value is the variant that corresponds to the outcome of the application of the variable rule (see Chapter 2 for a discussion of the relationships between variants). For example, for (t/d)-deletion, since we view the null variant as the outcome of the variable rule, we enter "0" (i.e. deletion) as the application value. GoldVarb also generates a cell file (*.cel), which it uses to calculate the results.

GoldVarb writes the output of its analysis to a results file (*.res), an example of which is shown in Figure 4.7. The results file displays the distribution (numbers) of tokens for all of the factors in each factor group, as well as the rates for each factor (also referred to as "marginals"). "Apps" indicates the number of tokens and rates of the variant we selected as the application value, while "Non-apps" are the number of tokens and rates of the other variant(s). Note that, because the first factor group is the dependent variable, what was originally coded as the second factor group is now the first factor group in the results.

4.3 Multivariate Analysis

To determine the statistical significance for each factor group, we could use the distribution of tokens shown in the results file as the observed



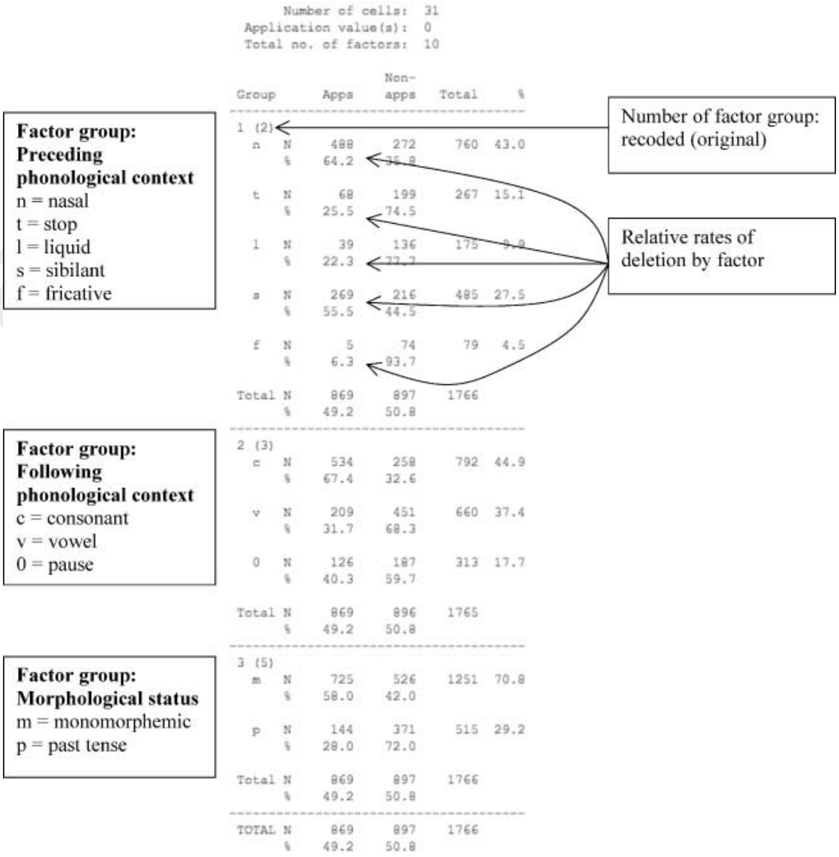
The application values are the values of the dependent variable (i.e. the first group in the condition file) to be used for creating cells.

If you specify only one value, then this will be the application value, and all other factors in the group will be counted as non-applications.

Example: Suppose the dependent variable has factors "abcd" (after recoding). Consider the following choices of application value(s):
- "a": binomial case; a = application; b, c & d = non-application.
- "ab": binomial; a = application; b = non-app; c & d omitted.
- "abc": trinomial case; d omitted.

Enter 1..9 factors here: 0

Figure 4.6 Application value window in GoldVarb.



```
          Number of cells:  31
     Application value(s):  0
     Total no. of factors:  10

                         Non-
Group        Apps        apps    Total      %
-------------------------------------------------
1 (2)
   n    N     488         272      760   43.0
        %    64.2        35.8

   t    N      68         199      267   15.1
        %    25.5        74.5

   l    N      39         136      175    9.9
        %    22.3        77.7

   s    N     269         216      485   27.5
        %    55.5        44.5

   f    N       5          74       79    4.5
        %     6.3        93.7

Total   N     869         897     1766
        %    49.2        50.8
-------------------------------------------------
2 (3)
   c    N     534         258      792   44.9
        %    67.4        32.6

   v    N     209         451      660   37.4
        %    31.7        68.3

   0    N     126         187      313   17.7
        %    40.3        59.7

Total   N     869         896     1765
        %    49.2        50.8
-------------------------------------------------
3 (5)
   m    N     725         526     1251   70.8
        %    58.0        42.0

   p    N     144         371      515   29.2
        %    28.0        72.0

Total   N     869         897     1766
        %    49.2        50.8
-------------------------------------------------
TOTAL   N     869         897     1766
        %    49.2        50.8
```

Factor group: Preceding phonological context
n = nasal
t = stop
l = liquid
s = sibilant
f = fricative

Factor group: Following phonological context
c = consonant
v = vowel
0 = pause

Factor group: Morphological status
m = monomorphemic
p = past tense

Number of factor group: recoded (original)

Relative rates of deletion by factor

Figure 4.7 Example of a results file for (t/d)-deletion.

frequencies for a chi-square test, as in Chapter 3 (though GoldVarb does not perform this test). However, as we noted above, we normally have a number of hypotheses about which factors are affecting the data, and factors may influence each other in various ways. For these reasons, we need recourse to multivariate analysis, which assesses the individual relative contribution of each factor to the observed variation when all factors are considered simultaneously.

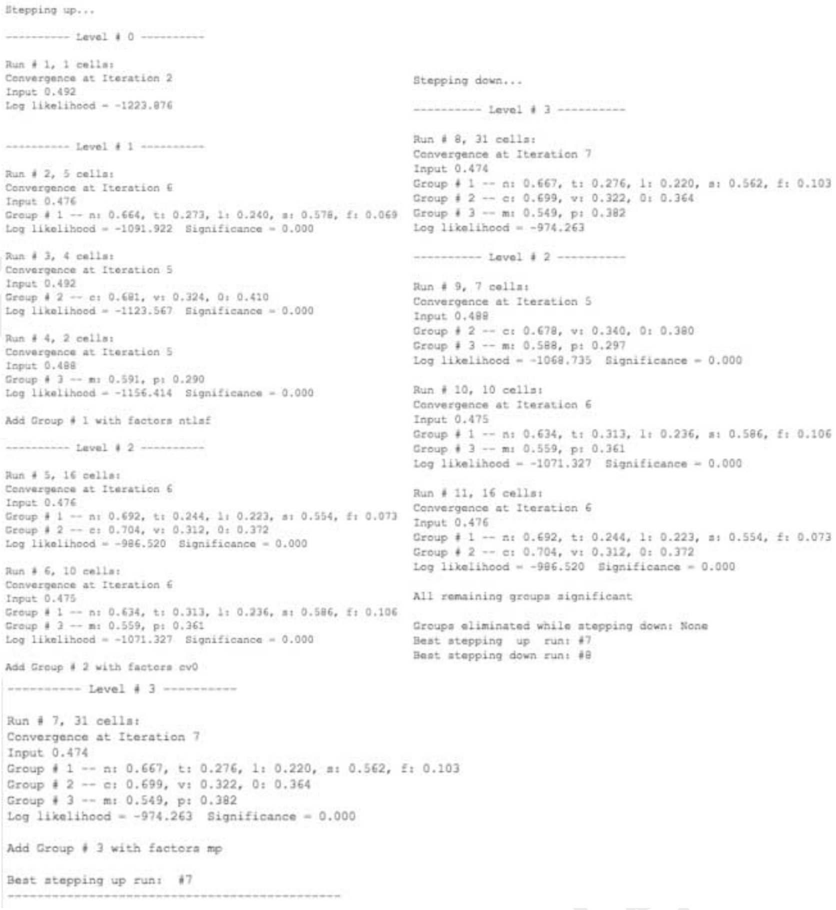
GoldVarb makes use of a type of multivariate analysis known as logistic regression. We will not discuss the specifics of the mathematics behind this procedure here (interested readers should consult Sankoff 1988a, Paolillo 2002 and Baayen 2008), except to note that GoldVarb only performs binomial multivariate analysis, where there is a choice of two results: application or non-application.[5] If there are only two variants, it does not greatly matter which is chosen as the application value, since the results for the non-application can be derived by subtracting each of the factor weights from 1. If there are more than two variants, we have to think about the relationships between the variants (see Chapter 3).

For each factor group, GoldVarb estimates the relative contribution (factor weight) that each factor in every factor group makes to the occurrence of the application value. Factor weights are centered on .5, such that anything above .5 favors the application value, while anything below .5 disfavors application. GoldVarb compares the expected distribution predicted by these estimates to the observed distribution and calculates the variance, or the distance between the expected and observed distributions. As a measurement of variance, instead of a chi-square value, GoldVarb uses a value of log likelihood, a measurement of how well the model fits the data. To determine which factor groups exert a statistically significant effect, GoldVarb performs a step-up/step-down procedure. In contrast with other statistical procedures that are concerned with determining how much of the variation is accounted for by the analysis, GoldVarb's procedure looks for the configuration of factors that provides the best fit to the observed distribution of variants.

In the step-up procedure, GoldVarb begins by calculating the overall probability that the "rule" applies (the input, also called the corrected mean) and the overall log likelihood, to provide a baseline of how well the overall probability predicts the distribution of data. At the first step-up, it adds each factor group to the input in turn and sees whether adding any of these factor groups improves the prediction of the model (improves the log likelihood) in a statistically significant way. If it does, that factor group is retained for the analysis. At the next step, it keeps that factor group and the input and again adds each of the remaining factor groups to the analysis in turn, seeing whether there are statistically significant improvements. It continues to step up until adding all factor groups produces no further improvement in the log likelihood, at which point it indicates the best step-up run.

The step-down procedure, which provides a check on the step-up procedure, begins by forcing all of the factor groups and the input into one analysis and calculating the log likelihood. At the first step-down, it takes away each factor group in turn and determines whether subtracting any factor group produces a statistically significant change in log likelihood. If so, that factor group is rejected. At the next step, it excludes that factor group and tries removing each of the other factor groups in turn, again checking for significant improvements in log likelihood. It continues to step down until taking away factor groups produces no improvement in the log likelihood, at which point it chooses the best step-down run. Figure 4.8 shows a step-up / step-down procedure for (t/d)-deletion.

The best step-up run and the best step-down run contain the same factor groups, those that are selected by GoldVarb as significantly improving the fit of the model to the data, and these are the factor weights



```
Stepping up...

---------- Level # 0 ----------

Run # 1, 1 cells:
Convergence at Iteration 2
Input 0.492
Log likelihood = -1223.876


---------- Level # 1 ----------

Run # 2, 5 cells:
Convergence at Iteration 6
Input 0.476
Group # 1 -- n: 0.664, t: 0.273, l: 0.240, s: 0.578, f: 0.069
Log likelihood = -1091.922  Significance = 0.000

Run # 3, 4 cells:
Convergence at Iteration 5
Input 0.492
Group # 2 -- c: 0.681, v: 0.324, 0: 0.410
Log likelihood = -1123.567  Significance = 0.000

Run # 4, 2 cells:
Convergence at Iteration 5
Input 0.488
Group # 3 -- m: 0.591, p: 0.290
Log likelihood = -1156.414  Significance = 0.000

Add Group # 1 with factors ntlsf

---------- Level # 2 ----------

Run # 5, 16 cells:
Convergence at Iteration 6
Input 0.476
Group # 1 -- n: 0.692, t: 0.244, l: 0.223, s: 0.554, f: 0.073
Group # 2 -- c: 0.704, v: 0.312, 0: 0.372
Log likelihood = -986.520  Significance = 0.000

Run # 6, 10 cells:
Convergence at Iteration 6
Input 0.475
Group # 1 -- n: 0.634, t: 0.313, l: 0.236, s: 0.586, f: 0.106
Group # 3 -- m: 0.559, p: 0.361
Log likelihood = -1071.327  Significance = 0.000

Add Group # 2 with factors cv0
---------- Level # 3 ----------

Run # 7, 31 cells:
Convergence at Iteration 7
Input 0.474
Group # 1 -- n: 0.667, t: 0.276, l: 0.220, s: 0.562, f: 0.103
Group # 2 -- c: 0.699, v: 0.322, 0: 0.364
Group # 3 -- m: 0.549, p: 0.382
Log likelihood = -974.263  Significance = 0.000

Add Group # 3 with factors mp

Best stepping up run:  #7
---------------------------------------------
```

```
Stepping down...

---------- Level # 3 ----------

Run # 8, 31 cells:
Convergence at Iteration 7
Input 0.474
Group # 1 -- n: 0.667, t: 0.276, l: 0.220, s: 0.562, f: 0.103
Group # 2 -- c: 0.699, v: 0.322, 0: 0.364
Group # 3 -- m: 0.549, p: 0.382
Log likelihood = -974.263

---------- Level # 2 ----------

Run # 9, 7 cells:
Convergence at Iteration 5
Input 0.488
Group # 2 -- c: 0.678, v: 0.340, 0: 0.380
Group # 3 -- m: 0.588, p: 0.297
Log likelihood = -1068.735  Significance = 0.000

Run # 10, 10 cells:
Convergence at Iteration 6
Input 0.475
Group # 1 -- n: 0.634, t: 0.313, l: 0.236, s: 0.586, f: 0.106
Group # 3 -- m: 0.559, p: 0.361
Log likelihood = -1071.327  Significance = 0.000

Run # 11, 16 cells:
Convergence at Iteration 6
Input 0.476
Group # 1 -- n: 0.692, t: 0.244, l: 0.223, s: 0.554, f: 0.073
Group # 2 -- c: 0.704, v: 0.312, 0: 0.372
Log likelihood = -986.520  Significance = 0.000

All remaining groups significant

Groups eliminated while stepping down: None
Best stepping  up  run: #7
Best stepping down run: #8
```

Figure 4.8 Step-up and step-down for (t/d)-deletion.

that we report. For example, (4.3) shows the best stepping-down run from the variable rule analysis of (t/d)-deletion in Figure 4.8. The input value (that is, the overall probability of deletion to occur in these data) is .474. All three factor groups are selected as significant: Group #1 (preceding phonological context), Group #2 (following phonological context), and Group #3 (morphological status). The decimal values within each factor group indicate the factor weight for each factor. For example, in Group #1, preceding nasals (n) have a factor weight of 0.667, preceding stops (t) have a factor weight of 0.276, and so on.

(4.3)
```
Run # 8, 31 cells:
Convergence at Iteration 7
Input 0.474
```

```
Group # 1—n: 0.667, t: 0.276, l: 0.220, s: 0.562,
  f: 0.103
Group # 2—c: 0.699, v: 0.322, 0: 0.364
Group # 3—m: 0.549, p: 0.382
Log likelihood = −974.263
```

It is good practice to report all of the factor groups included in the run, as well as indicating factor groups that were included in the run but were not selected as significant. Table 4.1 reports the results obtained in (4.3). Here, factor weights are rounded to the second decimal place (for readability) and are supplemented with the percentages and total number of tokens for each factor (obtained from the results generated before the step-up/step-down procedure in Figure 4.7). The relative strength of each factor group within the run is indicated by the range values,[6] which are obtained by subtracting the largest factor weight from the smallest factor weight in each factor group.

4.4 Identifying and Overcoming Interaction

As we mentioned above, GoldVarb does not require that data be distributed equally across all factors and factor groups (as ANOVA does), but the multivariate procedure used in GoldVarb does assume that all factor groups operate independently of each other. If they do not, the results generated by the program become questionable. One limitation of GoldVarb is that it does not identify dependence or interaction between factor

Table 4.1 Linguistic factors contributing to (t/d)-deletion in Toronto English.

	Total N:	1,776			
	Input:	.474			
				%	N
Preceding Phonological Context					
Nasal		.67		64	760
Sibilant		.56		56	485
Stop		.28		26	267
Liquid		.22		22	175
Fricative		.10		6	79
	Range:		57		
Following Phonological Context					
Consonant		.70		67	792
Pause		.36		40	313
Vowel		.32		32	660
	Range:		38		
Morphological Status					
Non-past		.55		58	1251
Past		.38		28	515
	Range:		17		

Factors not selected as significant: None.

groups, which may occur because of poor coding decisions (i.e. two factor groups are really testing the same thing twice), facts about language that cannot be avoided (see below), or sparse distribution of data.

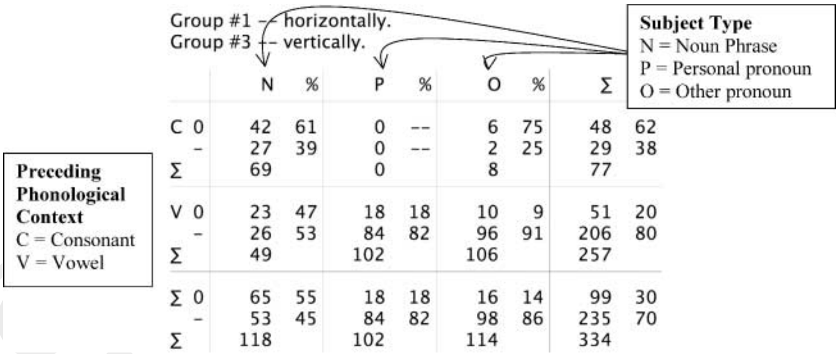
There are several indications that interaction is present: different factor groups are selected in the best step-up and step-down runs; warning messages appear during the step-up or step-down; the relative ordering of factor weights and percentages within a factor group do not match. As an example, consider Table 4.2, which shows a variable rule analysis of zero copula in third person singular contexts in African Nova Scotian English (adapted from Walker 2000b; see Chapter 6 for a more detailed discussion of this variable), examining the relative effects of the preceding and following phonological contexts, the grammatical category following the copula, and the type of subject. As the crossed lines indicate, the ordering of factors within the type of subject differs between the factor weights and the percentages, suggesting interaction of this factor group with another factor group.

The best way to identify the source of interaction is to cross-tabulate each factor group against every other factor group and look for sparse distribution or gaps in the distribution of data.[7] Figure 4.9 shows a cross-tabulation of two of the factor groups analyzed in Table 4.2. Concentrating on the Σ values, which indicate the total number of tokens in each cell, we see that there is one cell with no data: the combination of personal pronouns and preceding consonants (which corresponds to a

Table 4.2 Factors contributing to zero copula in third person singular contexts in African Nova Scotian English (adapted from Walker 2000b).

	Total N:	334			
	Input:	.247			
				%	N
Following Grammatical Category					
V-*ing*		.78		56	61
gonna		.66		54	26
Adjective		.59		31	64
Locative		.58		40	38
NP		.28		11	145
	Range:		*50*		
Type of Subject					
NP		.61		55	118
Other Pronoun (*it*, *what*, *that* . . .)		.54		14	114
Personal Pronoun (*he*, *she*)		.33		18	102
	Range:		*28*		
Preceding Phonological Context					
Consonant		.68		62	77
Vowel		.44		20	257
	Range:		*24*		

Factor groups not selected as significant: Following Phonological Context.



Group #1 -- horizontally.
Group #3 -- vertically.

Subject Type
N = Noun Phrase
P = Personal pronoun
O = Other pronoun

Preceding Phonological Context
C = Consonant
V = Vowel

		N	%	P	%	O	%	Σ	
C	0	42	61	0	--	6	75	48	62
	-	27	39	0	--	2	25	29	38
Σ		69		0		8		77	
V	0	23	47	18	18	10	9	51	20
	-	26	53	84	82	96	91	206	80
Σ		49		102		106		257	
Σ	0	65	55	18	18	16	14	99	30
	-	53	45	84	82	98	86	235	70
Σ		118		102		114		334	

Figure 4.9 Cross-tabulation of subject type against preceding phonological context for zero copula in African Nova Scotian English (Walker 2000b).

consonant-final subject preceding the copula). In other words, no personal pronouns end in a consonant. This interaction is therefore due to a fact of the English language.

Interaction may be overcome in a number of ways. If we have defined factor groups poorly or we have made poor coding decisions, we must return to the token file. Combining factors within factor groups is done within the condition file (using the "recode" function). This option is best for sparse distribution of data, though combining factors should be based on linguistic principles. For example, we could overcome the interaction shown in Figure 4.9 by combining the two pronoun factor groups (personal pronouns and other pronouns) into one (pronouns). We can also use condition files to combine two or more factor groups into a single "interaction group". For example, to work around the interaction shown in Figure 4.9, we could combine the preceding phonological context and subject type into one factor group, using the AND function, to create a new factor group of five factors: personal pronouns (which always end in a vowel), other pronouns ending in a vowel, other pronouns ending in a consonant, NP subjects ending in a vowel, and NP subjects ending in a consonant. If neither option is feasible, we can make use of log likelihood to decide between competing analyses. Performing separate multivariate analyses, each of which excludes one of the interacting factor groups in turn, we compare the log likelihoods of the best runs for these analyses, with the lowest log likelihood value (i.e. that closer to zero) indicating which analysis provides the best fit of the model to the data.[8] We will see examples of this type of comparative analysis in later chapters.

4.5 Conclusion

In this chapter, we discussed multivariate analysis with the program GoldVarb, the most common program for variationist analysis. We began by discussing techniques of coding and formatting the data for use in GoldVarb. We discussed the use of condition files in reconfiguring and analyzing the token file. We discussed how to generate percentages for each factor and the need for determining statistical significance and for considering the contribution of each factor when all factors are considered together. This led to a discussion of performing multivariate analysis using the step-up/step-down procedure. We also touched on some of the limitations of GoldVarb, specifically its inability to identify and deal with interaction or overlap between factor groups. We suggested ways of overcoming this limitation, how to identify interaction using cross-tabulation and how to overcome interaction by reconfiguring the factor groups in different ways.

Having established the basic concepts of variation and variables and provided a thorough discussion of the analytical and statistical techniques, in the following two chapters we will deal in more detail with the types of variation that exist at different levels of the linguistic system. A common theme of these chapters is the variationist concern to detail not only the presence of a particular feature, but also its overall rate of occurrence and, more importantly, the conditioning of the variation by features of the linguistic context.

4.6 Further Reading

Baayen, Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Bayley, Robert. 2002. The quantitative paradigm. In J.K. Chambers, P. Trudgill & N. Schilling-Estes (eds.), *The Handbook of Language Variation and Change*. Oxford: Blackwell, 117–41.

Guy, Gregory R. 1988. Advanced VARBRUL analysis. In K. Ferrara, B. Brown, K. Walters & J. Baugh (eds.), *Linguistic Change and Contact*. Austin: Department of Linguistics, University of Texas at Austin, 124–36.

Paolillo, John C. 2002. *Analyzing Linguistic Variation: Statistical Models and Methods*. Stanford: CSLI Publications.

Tagliamonte, Sali. 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Young, Richard & Robert Bayley. 1996. VARBRUL analysis for second language acquisition research. In D. Preston (ed.), *Second Language Acquisition and Linguistic Variation*. Amsterdam/Philadelphia: Benjamins, 253–306.

5 Variation in Sound Systems

5.0 Introduction

In the previous chapters, we laid the groundwork for topics that will be explored in the rest of the book. We defined the concept of variation in language and introduced the methodological constructs of the variable and the variable context. We also outlined the methodological and statistical procedures involved in the analysis of variation and multivariate analysis using GoldVarb. In this chapter and the next, we examine variation at different levels of the linguistic system, making a broad division between variation at the level of sound systems (i.e. phonetic and phonological variation) and variation at the level of grammatical systems (here including morphological, syntactic and discourse variation). As we will see, this division is a bit artificial, since some variables cut across sound systems and grammatical systems. However, the variables in each of these two areas are concerned with slightly different research questions and methods.

The first type of variable to be studied quantitatively, and arguably still the most commonly studied, involves variation at the level of sound systems. For this reason, many of the methodological and theoretical issues have been discussed extensively. However, there remain a number of unresolved or controversial issues even for these variables. In addition, most studies of variation in sound systems focus on correlating the variation with social factors such as social class, age, sex/gender and ethnicity. Less attention has been paid to the conditioning of variation in sound systems by language-internal factors.

Where do we draw the line between "phonetic" and "phonological" variation, or does such a line even exist? All variation is ultimately phonetic, since one variant is (or is not) pronounced. The question is whether the variation takes place at the level of speech production (phonetics) or whether it is a part of the sound system of the language (phonology). We might avoid this question by referring to such variation as "phonetic or phonological", but this phrase becomes cumbersome with repetition. For this reason, in the remainder of the book we will use the term "phonic" to refer to variation that may occur at the level of phonetics or phonology.

This chapter explores the issues involved in the analysis of phonic variation. We begin by examining types of phonic variation and the differences between variable and categorical processes in sound systems, before proceeding to issues in defining the variable context for phonic variables, dividing such variables into their variant realizations, and calculating overall rates of occurrence. We then discuss the different factors hypothesized to condition variation in sound systems, including not only factors having to do with the phonetic or phonological context, but also syntactic, lexical and discourse factors.

5.1 Types of Phonic Variable

A survey of the literature on phonic variation reveals a wide array of variable processes that occur at different levels of the sound system. Consonants may be deleted (5.1) or shifted in place (5.2) or manner of articulation (5.3), among other processes.

(5.1) (t/d)-deletion (English)
west ~ *wes'*
(5.2) (ing) (English)
singing (velar) ~ *singin'* (coronal)
(s)-aspiration (Spanish)
entonces "then" (dental) ~ *entonceh* (laryngeal)
(5.3) (th)-stopping (English)
them (fricative) ~ *dem* (stop)

Vowels may be deleted (5.4), raised or lowered (5.5), fronted or backed (5.6), diphthongized or monophthongized (5.7).

(5.4) (u)-deletion (European Portuguese; Silva 1994)
gatu "cat" ~ *gat_*
(5.5) (ay)-raising (Canadian English; Chambers 1973)
right: [rajt] ~ [rəjt]
Lax vowel lowering (Canadian English; Clarke et al. 1995)
pen: [pɛn] ~ [pæn]
(5.6) (uw)-fronting (Californian English)
move: [muwv] ~ [mʉwv]
(æ)-retraction (Canadian English; Clarke et al. 1995)
pat: [pæt] ~ [pat]
(5.6) Canadian French Diphthongization
tête "head": [tɛt] ~ [tajt]
(5.7) (ay)-monophthongization (Southern US English)
white: [wajt] ~ [wa:t]

If we compare phonic variables, such as those illustrated in (5.1) to (5.7), with categorical rules or processes in phonology (see Chapter 2), we may be struck more by the similarities than by the differences. In fact, for any variable process, it is not difficult to find a corresponding categorical process. Thus, the only difference between the two is one of degree rather than kind. Indeed, as we will see below, we might view categorical processes as variable processes with a very high probability of application.

5.1.1 Defining the Variable Context

As we discussed in Chapter 2, an important step in the analysis of variation (perhaps the most important step) is defining the variable context, which answers the question of where the speaker has a choice between forms. Defining the variable context for phonic variables seems relatively straightforward, certainly more so than for grammatical variation (see Chapter 6). However, before conducting an analysis of phonic variation, there remain a number of important questions that need to be addressed.

Since we normally define the variable context for phonic variables in terms of a phonetic or phonological context, as such it constitutes a linguistic analysis. Specifically, we are making a statement of our assumptions about the nature of the underlying forms of the lexical items included in that context. For example, when we define the variable context for (t/d)-deletion as “every word-final /t/ and /d/ in a consonant cluster”, we assume that we know which lexical items contain an underlying word-final /t/ or /d/. This assumption is unproblematic for standard varieties of English, but may not hold across all varieties. For example, in listening to recordings of African Nova Scotian English, I heard many speakers produce a plural form *ghostes* [gosəz], instead of *ghosts* [gosts], suggesting that for these speakers the underlying form of the word *ghost* has no underlying final /t/.

Defining the variable context for vowels raises even more problems. Given the variability in vowel production more generally, how do we know which words contain instances of the vowel of interest to us? One approach is to define the variable context on the basis of the phoneme. For example, in his study of New York City English, Labov (1966) defined the variable context of (oh) as all words containing the phoneme /ɔ/: *caught*, *coffee*, *dog*, etc. This approach is useful for studying variation within a linguistic variety, but may be less useful for cross-variety comparison, since not all varieties of English make the same phonemic distinctions among vowels, nor do they consistently include the same lexical items in each class of vowel phoneme. For example, in Canadian English, /ɔ/ and /ɑ/ are not distinct, so that *cot* and *caught* are pronounced the same.

An alternative approach, promoted by Wells (1982) and commonly used in variationist research in the UK, defines variables in terms of word-classes rather than phonemes. Table 5.1 shows the list of word-classes

Table 5.1 Standard lexical sets for English vowels (adapted from Wells 1982: xviii–xix)

Class	Phoneme		Examples
	RP	GenAm	
KIT	ɪ	ɪ	*sick, bridge, milk, busy . . .*
DRESS	e	ɛ	*step, edge, friend, ready . . .*
TRAP	æ	æ	*tap, badge, hand, cancel . . .*
LOT	ɒ	ɑ	*stop, dodge, possible, quality . . .*
STRUT	ʌ	ʌ	*cup, budge, trunk, blood . . .*
FOOT	ʊ	ʊ	*put, full, good, look . . .*
BATH	ɑː	æ	*brass, ask, dance, calf . . .*
CLOTH	ɒ	ɔ	*cough, cross, long . . .*
NURSE	ɜ	ɜr	*hurt, burst, jerk, term . . .*
FLEECE	iː	i	*speak, leave, feel, people . . .*
FACE	eɪ	eɪ	*tape, cake, veil, day . . .*
PALM	ɑː	ɑ	*father, spa, psalm . . .*
THOUGHT	ɔː	ɔ	*sauce, hawk, jaw, broad . . .*
GOAT	əʊ	o	*soap, joke, home, know . . .*
GOOSE	uː	u	*shoot, mute, huge, view . . .*
PRICE	aɪ	aɪ	*write, arrive, high, buy . . .*
CHOICE	ɔɪ	ɔɪ	*noise, join, toy, royal . . .*
MOUTH	aʊ	aʊ	*out, house, loud, cow . . .*
NEAR	ɪə	ɪ(r	*beer, fear, beard . . .*
SQUARE	ɛə	ɛ(r	*care, where, scarce, vary . . .*
START	ɑː	ɑ(r	*far, sharp, farm, heart . . .*
NORTH	ɔː	ɔ(r	*for, war, short, warm . . .*
FORCE	ɔː	o(r	*four, wore, porch, story . . .*
CURE	ʊə	ʊ(r	*poor, tourist, pure . . .*

proposed by Wells, along with the corresponding phonemes in British Received Pronunciation (RP) and General American (GenAm). This approach has the advantage of providing for cross-variety comparison, because the number of word-classes exceeds the number of phonemic distinctions that exist in any variety of English. For example, it can be noted for Canadian English that LOT and THOUGHT are combined as a single class, whereas they must be separated for New York City English. An added value of this system is that it is more mnemonic than the phonemic approach, in that variables are more readily identifiable through the example words used to name each class than are the symbols used in phonemic approaches, which may differ across varieties of English. The major problem with the word-class approach is that it is not always easy to determine which class each lexical item belongs to, especially in varieties whose vowel systems have not already been analyzed.

In defining the variable context, we need to specify not only which forms to include as tokens, but also which forms to exclude. An important environment to exclude is neutralization contexts, in which it is difficult or impossible to determine which variant is realized. For (t/d)-deletion, we normally exclude following coronal stops (such as *west Toronto*), as well as following interdentals (e.g. *they were sold there*). In both cases, co-articulation with the following context makes it difficult to tell whether the /t/ or /d/ has been deleted. Similarly, for (s)-deletion in Spanish, following sibilants are excluded (e.g. *Quieres salir?* "Do you want to leave?"). Some claim to be able to distinguish the variants even in such contexts, but since not all researchers are so blessed, such exclusions need to be applied consistently to ensure comparability. For vowel variables, unstressed syllables are normally excluded, since reduced vowels often neutralize phonemic distinctions. For example, unstressed vowels in English generally neutralize to [ə] or [ɪ], making it impossible to distinguish among variants.

5.1.2 Variants and Relative Frequencies

Up until this point, we have been acting as if all variables are nominal: that is, each variant can be assigned to discrete category. In fact, VARBRUL analysis assumes that variables are nominal. However, as anyone who has worked on naturally-occurring speech knows, phonetic production varies along a number of dimensions, and a sound is rarely pronounced exactly the same way every time. For example, although we have been distinguishing two variants for (t/d)-deletion (pronounced or not pronounced), in reality there are different degrees of closure and release. One option is to make finer distinctions among variants (e.g. fully pronounced, partially pronounced, deleted), another is to consider any audible pronunciation as non-deletion, and yet another is simply to exclude any intermediate forms.

These options are more difficult to employ in the case of truly continuous variables, which lie along a continuum of realizations. For example, vowel variants range along dimensions in the phonetic space, such as low/high and front/back (often simultaneously). Continuous variables raise problems for variationist analysis, not only in terms of defining the variants, but also in terms of calculating relative frequencies. We will consider two methods for studying continuous variables: auditory or impressionistic analysis, and acoustic or instrumental analysis.

Auditory or Impressionistic Analysis

One option for dividing a continuous variable into variants is to divide the articulatory continuum into intervals (in effect, to turn a continuous variable into a nominal variable) on the basis of auditory differences. For example, in Labov's (1963) study of English in Martha's Vineyard, Massachusetts, the onset of the diphthongs /ay/ (as in *light*, *ice*) and /aw/ (as in *house*, *out*) varied in height between [a] and [ə] (Figure 5.1). Based on differences he could discern by listening (auditory or impressionistic

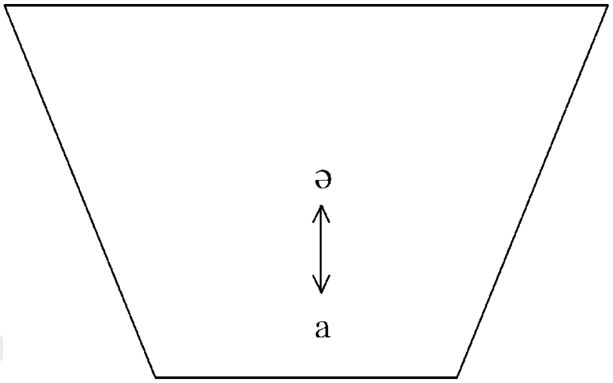


Figure 5.1 Centralization of (ay) and (aw) on Martha's Vineyard (Labov 1963).

classification), he divided the vowel continuum into a scale of six variants, ranging from the most standard [a] to the most non-standard [ə], as shown in Scale I in Table 5.2. After comparing his impressions with instrumental measurements (see the next section), he revised the six-point scale to four points (Scale II).

Although dividing the vowel continuum into intervals allows us to distinguish among variants, it raises issues in calculating frequencies. In the Martha's Vineyard example, there are four variants, which means we must consider the possibility of multiple (possibly ordered) variable rules, as illustrated in (5.8).

(5.8) a. [a] → [a̝]
b. [a̝] → [ɐ̞]
c. [ɐ̞] → [ɐ]

Table 5.2 Weighted index scores of (ay) and (aw) variables for one fisherman in Martha's Vineyard (adapted from Labov 1963).

			(ay) index		(aw) index	
Variant	Scale I	Scale II	# tokens	Weight	# tokens	Weight
[a]	1	0	49	0	32	0
[a̝]	2	1	27	27	10	10
[ɐ̞]	3					
[ɐ]	4	2	24	48	4	8
[ɐ̝]	5					
[ə]	6	3	0	0	0	0
		Total:	100	75	46	18
			$\frac{75}{100} \times 100=$		$\frac{18}{46} \times 100=$	
Weighted index score:			75		39	

d. [ɐ] → [ɐ̝]
e. [ɐ̝] → [ə]

However, it is easier to assume a single gradient rule, with varying degrees of application. This assumption calls for a method of calculation that allows us to express central tendencies with a single figure. Labov's solution was to use a weighted index score: each variant is assigned a relative weight (ranging from 0 to 3), the number of tokens of each variant is multiplied by its weight, all scores are totalled, and the total weight is divided by the total number of tokens and then multiplied by 100. A weighted index score of 0 would be the most standard, while a score of 100 would be the most non-standard. Table 5.2 shows the calculation of index scores for (ay) and (aw) for one speaker.

Although weighted index scores are useful in reducing a continuum of realizations to a single numerical value, they are not without problems. Dividing the continuum into intervals implies that each variant lies at an equal (social or phonetic) distance from its neighbours. However, it is not clear that impressionistic measures have that degree of accuracy or that such distances are perceptually important. More problematically, quite different distributions of tokens among the variants can result in similar index scores, which may obscure important differences between speakers.

Instrumental or Acoustic Analysis

Since Labov, Yaeger and Steiner (1972), studies of continuous variables have tended to make use of instrumental or acoustic analysis, using programs such as Praat (Boersma & Weenink 2008) and Plotnik (Labov 2008). For each vowel token, we take measurements (in Hz) of the first (F1) and second (F2) formant frequencies, which correspond to vowel height and frontness, respectively. Plotting F1 as the vertical axis and F2 as the horizontal axis provides a graphic representation that roughly approximates the positions of vowels in the human mouth. For example, as part of the Canadian Vowel Shift (Clarke et al. 1995), the front lax vowels /ɛ/ and /æ/ are variably shifted to phonetic values more like [æ] and [a], respectively (so that *pen* sounds like *pan* and *pan* sounds like *pawn*). Figure 5.2 plots the mean F1 and F2 values of shifted and non-shifted variants of the variables /ɛ/ (eh) and /æ/ (ah) for five groups of speakers in Toronto (adapted from Hoffman, in preparation). As this plot shows quite clearly, both (eh) and (ah) are variably retracted and/or lowered to different degrees in each group.

Instrumental measurement permits a more precise identification of the location of the vowel than does impressionistic coding, but abandoning nominal variants means that the quantitative procedures discussed in Chapter 3 are no longer appropriate. Specifically, we cannot make use of the multiple regression feature of GoldVarb, since it relies on nominal

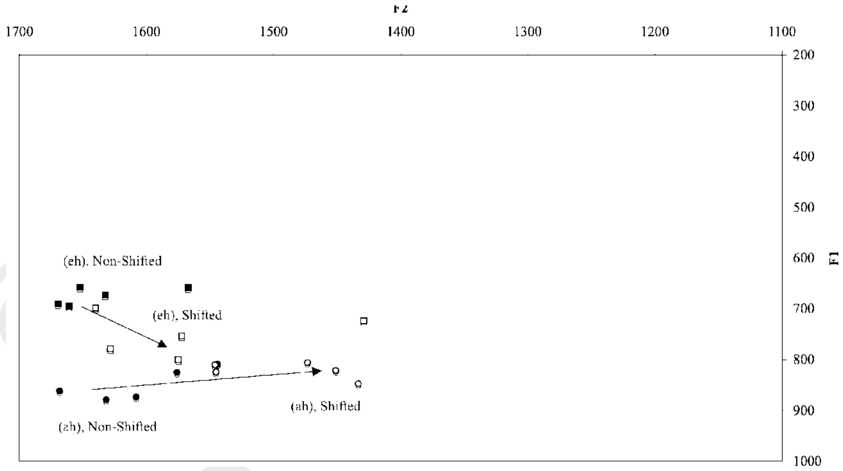


Figure 5.2 Shifted and non-shifted tokens of (eh) and (ah) in Toronto English (adapted from Hoffman, in preparation).

variants. Instead, we can correlate formant frequency values for F1 and F2 with each of the factors hypothesized to condition the variation, using measurements of linear correlation. However, analyzing correlations on a factor-by-factor basis raises the same problems of single-factor analysis of nominal variables discussed in Chapter 3. We can still perform multivariate analysis, but we must use methods other than those available in GoldVarb. Statistical programs such as SPSS and R incorporate multivariate analysis for linear values which provide measurements of statistical significance, as well as coefficients that measure the direction and strength of contribution that each factor makes to the (dependent) variable. As an example, consider Table 5.3, which shows a multiple-regression analysis of phonetic factors conditioning the variable raising of /æ/ in one speaker from Labov's study of Philadelphia. In Table 5.3, the

Table 5.3 Effect of phonetic features on the height of (æh) in the speech of Carol Meyers, Philadelphia (Labov 1994: 466)

	Coefficient	*t*	*p (2-tail)*
Constant	2104		
Following nasal	158	3.57	0.000
Following /d/ (*mad*, *bad*, *glad*)	238	2.44	0.016
Preceding nasal	99	1.22	0.225
Preceding obstruent plus /l/	–210	–2.72	0.007
Two following syllables	–410	–2.75	0.007
Secondary stress	–95	–1.95	0.053

N = 149 Multiple *r* = .47 Squared multiple *r* = .22 F = 6.62

procedure assigns each factor a coefficient value that indicates the strength and direction of its effect: a positive value indicates a favoring effect, while a negative value indicates a disfavoring effect. Statistical significance is determined through a two-tailed *t*-test, with *p*-values below .05 considered significant. In Table 5.3, we see that following and preceding nasals all have positive values and thus favor raising, while preceding obstruents plus /l/, two following syllables and secondary stress all inhibit raising. However, the *p*-values for preceding nasals and secondary stress are greater than .05 and thus are not significant.

In current research on vowel variables, instrumental analysis is the rule rather than the exception, though it is not without problems. Since the exact values of formant frequencies are determined by the size and shape of individual speakers' vocal tracts, values differ from speaker to speaker. Studies of vowel variables thus need to employ a normalization procedure to filter out these differences, just as human perception does. There are a number of methods of normalization available, mostly divided according to whether they are based on the information contained in a single vowel (vowel intrinsic) or based on the distribution of all vowels for each speaker (vowel extrinsic) (see Adank et al. 2004 for an overview). However, since there is no standard method of normalization, it is often difficult to compare results across studies of phonic variation. Moreover, it is a matter of controversy whether normalization procedures factor out potentially important or salient aspects of variation between and within speakers, or even whether they add further variation to the data. More generally, while F1 and F2 values provide a good indication of some components of vowel production, they may not capture all of the aspects of the speech signal that are perceived as salient by the listener. For these reasons, it is good practice to combine impressionistic and instrumental coding of vowel variables.

5.2 Conditioning of Phonic Variation

As we noted in Chapter 3, the analysis of linguistic variation is concerned not only with determining the overall frequency of each variant, but also with correlating differences in frequency with contextual factors, each of which represents a hypothesis and a potential explanation for the observed variability. We can divide these explanations into two broad types: those "external" to the linguistic system and those "internal". External explanations involve what we might call the socio-symbolic uses of variation: that is, the indication of membership in different social groups, such as social class, sex/gender, level of education, ethnicity, and so on. Obviously, a large proportion of phonic variation conveys these functions, and in fact most studies of phonic variation tend to concentrate on this type of conditioning. Also language-external, though less concerned with socio-symbolic function, are explanations having to do with physiological considerations, such as neurology, language production and processing, limitations of memory and attention, and so on. Without wishing to downplay external explanations (in fact, in many cases they may be more important than linguistic explanations), in the remainder of this chapter we will focus on language-internal explanations of phonic variation: lexical, phonic, grammatical.

5.2.1 Lexical Conditioning

Since the definition of the variable context for phonic variables entails assumptions about the underlying representation of the lexical context in which the variable occurs, any analysis of phonic variation must take into account potential lexical effects, due either to individual lexical items or to lexical classes. Even as early as Labov's (1963) Martha's Vineyard study, he noted that vowel centralization was more common with certain lexical items, though he did not investigate this effect quantitatively. In subsequent work, it has been standard to exclude certain lexical items as part of defining the variable context. For example, studies of (t/d)-deletion usually exclude words such as *and* and the negative contraction *-n't* because of their extremely highly frequency in speech and their almost categorical deletion (anyone who listens to a stretch of natural speech will become aware of just how infrequent the full form [ænd] is!). If we were to include such forms in the token file, not only would their frequency mean that they would occupy a large portion of the data, but their high rates of deletion would also inflate the overall rate. To control for possible unforeseen effects of other lexical items, it is also common practice to take no more than a few tokens of each lexical item (say, five) per speaker. Although studies of phonic variables have not explored the quantitative effects of individual lexical items in detail, lexical exceptions have been noted for both vowel and consonant variables. For example, Labov's (1994) study of short (aeh) in Philadelphia finds the words *mad*, *bad* and *glad* to constitute exceptions to the general inhibiting effect of words ending with voiced stops on the tensing of /æ/. Walker's (2008) study of (t/d)-deletion in Toronto English, shown in Table 5.4, shows that, even after restricting the number of tokens per lexical type per speaker, certain highly frequent lexical items, such as *different* and *went*, favor deletion at rates (87–88%) much higher than the average (44%).

5.2.2 Phonic Conditioning: Phonetics and Phonology

Most work on phonic variation has been more concerned with the conditioning by phonetic and phonological factors. What is the difference between phonetic and phonological conditioning? As we noted above, it may be difficult to draw a line between these two systems. Generally, phonetics concerns the production of speech, without necessarily taking

Table 5.4 Frequency of lexical type and rate of (t/d)-deletion in Toronto English (Walker 2008).

	N	% deletion
High Frequency:		
different	67	87
went	63	86
friend	60	55
first	51	61
old	45	16
around	44	55
want	38	79
told	30	53
most	30	47
worked	29	52
last	25	56
end	25	8
left	22	32
called	22	27
lived	21	10
Low Frequency	713	38
Medium Frequency	407	40
Total	1692	44

into account the organization of sounds within a particular language, while phonology concerns the organization of sounds within and across languages. If we accept this division, it makes predictions about different types of conditioning. Phonetic explanations of variation are predicted to stem from considerations having to do with speech production, while phonological explanations are predicted to arise from the organization of sounds within the individual language. More generally, we predict that phonetic explanations are (potentially) universal, while phonological explanations are language-specific. For example, Labov (1963) finds that raising of (aw) and (ay) in Martha's Vineyard is less frequent if the following sound is velar. This effect has an articulatory explanation: the raising of the back of the tongue to the velum inhibits the raising of the tongue centre in producing the vowel. Thus, to explain this effect we can appeal to universal properties of human articulation and we do not need to refer to the phonology of English.

In some cases, though, the distinction between phonetic and phonological conditioning may be difficult to disentangle. For example, many studies have found that a following consonant favors (t/d)-deletion over a following vowel, as shown in the left-hand side of Table 5.5. One explanation for this effect is articulatory: cross-linguistically, consonant-vowel syllables are the most common, so this effect may be driven by ease of articulation: deleting the [t] or [d] at the end of a syllable is one way of getting closer to CV syllable structure. However, a phonological

Table 5.5 Two one-level binomial analyses of the effect of following phonological context on (t/d)-deletion in Toronto English (N=2,251).

		%	N			%	N
Consonant	**.68**	66	1013	Stop*	**.84**	83	226
Pause	.39	37	380	Nasal*	**.78**	77	155
Vowel	.33	32	858	[l]*	**.73**	72	72
				Fricative*	**.68**	67	173
				[r]	**.66**	65	37
				[w]	**.61**	60	148
				[ʃ]/[ʒ]	.44	43	14
				[j]	.43	41	109
				[h]*	.41	39	79
				Pause	.38	37	380
				Vowel	.33	32	858
Log likelihood:	–1433.382			Log likelihood:	–1382.852		
df:	2			df:	9		

$\chi^2 = 101.06$, df = 7, $p < .05$ \ significant

* Does not form a possible onset with [t] or [d].

explanation is also possible: if a vowel follows, the [t] or [d] can resyllabify as the onset of the following syllable, thus avoiding deletion (because it is no longer in the coda of the preceding syllable). We can decide between these competing explanations by looking in more detail at the effect of the following consonant. If resyllabification is a cross-linguistic process, driven by a universal or articulatory preference for CV syllables, there should be no difference in effect between different types of following consonant. However, if the process is language-specific, constrained by the phonotactics of English, resyllabification (i.e. retention) should be preferred with following consonants that can form a possible onset with /t/ and /d/ in English.

In Table 5.5, we test these two predictions by analyzing the effect of following segment on (t/d)-deletion in Toronto English in two different ways. (In each case, the results show a one-level binomial step-up for the factors conditioning deletion.) First, if we use the simpler division of consonant/vowel/pause (the left-hand side of Table 5.5.), a following consonant favors deletion, while a following vowel or pause disfavor, as expected from previous studies. Examining the nature of the following consonant in finer detail (the right-hand side of Table 5.5) shows that those following segments which do not form a possible onset with /t/ or / d/ in English (stops, nasals, [l]) have a higher rate of deletion than those which may form an onset ([r], [w], [ʃ]/[ʒ]). A chi-square test comparing the log likelihoods for the two analyses shows that the finer breakdown of following consonant provides a significantly better fit to the data. These results provide some evidence for phonological, as opposed to phonetic, conditioning.

Other cases of language-internal conditioning are more clearly phonological. For example, the phonological context preceding the /t/ or /d/ has been found to be significant for (t/d)-deletion. Studies have tried to account for these effects by appealing to relative ease of articulation or the sonority hierarchy, both of which are more universal (i.e. phonetic) explanations. Guy and Boberg (1997) suggest that the effects of the preceding segment could be explained with reference to phonological features: specifically, the more features shared between the preceding segment and /t/ or /d/, the more likely is deletion. They test this hypothesis by dividing the nature of the preceding segment not in terms of sound-classes (such as stops, sibilants, etc.) but rather in terms of the number of features shared, as shown in the left-hand side of Table 5.6. As predicted, preceding segments that share two features (stops, sibilants and /n/) favor deletion more highly than preceding segments that share only one feature (laterals, non-coronals and nasals). Further, they question whether each feature ([±coronal], [±continuant] and [±sonorant]) has the same effect and tested this hypothesis by running an analysis in which each feature constitutes a separate factor group, as shown in the right-hand side of Table 5.6. Note that while each feature exerts an independent statistically

Table 5.6 Two independent variable rule analyses of factors contributing to (t/d)-deletion in Philadelphia English (Guy & Boberg 1997) (N=1,071).

Preceding Segment			**Sonority**			
/t,d/	[+cor, −son, −cont]	KO	[−sonorant]		.58	
/s,z,ʃ,ʒ/	[+cor, −son]	.69	[+sonorant]		.42	
/p,b,k,g/	[−son, −cont]	.69		*Range:*		*16*
/n/	[+cor, −cont]	.73	**Continuancy**			
/f,v/	[−son]	.55	[−continuant]		.65	
/l/	[+cor]	.45	[+continuant]		.35	
/m,ŋ/	[−cont]	.33		*Range:*		*30*
/r/	?	.13	**Coronal Place**			
			[+coronal]		.65	
			[−coronal]		.35	
				Range:		*30*
			Voice (preceding obstruents)			
			[α voice]		.64	
			[−α voice]		.36	
				Range:		*28*
Log likelihood = −533.173			Log likelihood = −535.033			

χ^2=3.72, df=3, p>.25 ∴ not significant

significant effect on deletion, the effect of [–sonorant] is weaker (the range is half that of the other factor groups). Since the difference between the log likelihoods of the two analyses is not significant (p>.25), the effects observed on the left-hand side of the table can be explained by those on the right-hand side.

Other elements of the phonic context besides adjacent segments have been explored as conditioning factors. One such area is suprasegmental features, such as tone, intonation, stress and prosody. Walker (1995) examines segmental and suprasegmental factors conditioning postvocalic (r)-deletion in African Nova Scotian English (5.9). In addition to examining the preceding context (vowel nucleus) and the following context, he also examined whether the syllable in which the /r/ occurred received primary stress (5.9a–b) or non-primary stress (secondary stress or unstressed) (5.9c–d).

(5.9) a. all kinds of cards [kɑ:dz] (AN11/042)
b. for three o'clock service [ˈsɹ̩ˌvɪs] (AN31/087)
c. from the graveyard [ˈgreʲvˌyɐɹd] hill (AN70/085)
d. for the sugar [ˈʃʊˌgə] (AN79/186)

As shown in Table 5.7, only the combined factor group of vowel nucleus and stress was selected as significant. There is a tendency for non-front vowels ([a] and [o]) to favor deletion, but the effect of stress does not operate in the same direction for all nuclei. For most nuclei, primary stress is likely to lead to more deletion than non-primary stress. For syllabic [r] (5.9b, 5.9d), primary stress is likely to lead to less deletion than non-primary stress. This result suggests that the two sets of nuclei are not phonologically equivalent, a result that Walker uses to argue that (r)-deletion should be not be treated as the same rule in both contexts.

Table 5.7 Factors contributing to postvocalic (r)-deletion in African Nova Scotian English (Walker 1995) (N=650).

		Syllable Stress	
	Primary		**Non-primary**
Nucleus:			
[r]	.12	<	.60
[a]	.74	>	.55
[e]	.43	>	.27
[i]	.37		—
[o]	.68	>	.50
[aw]	—		.43
[ay]	.43		—

Factors not selected: Following environment; Grammatical category; Edge of prosodic word.

5.2.3 Grammatical Conditioning: Morphology, Syntax, Discourse

We have seen evidence that phonic variables may be conditioned by elements of the phonetic or phonological context, whether adjacent segments or properties of the suprasegmental context. In this section, we will see evidence that phonic variables may also be conditioned by elements of the grammatical context, such as their morphological status or the morphological structure of the word in which they occur, their syntactic position or their function in discourse.

The morphological status of the phonic variable and the morphological structure of the word in which it occurs have been found to have an effect in a number of studies (see Guy 1980). English (t/d)-deletion is more frequent if the [t] or [d] is part of the preceding morpheme ("monomorphemic"), as in *mist*, than if it is a separate morpheme marking past tense, as in *missed*. One explanation for this effect is "functional": since the [t] or [d] serves to mark a morphological distinction (tense) in past-tense forms, deletion would lead to a loss of information and potential ambiguity, and thus tends to be avoided. In contrast, deleting a monomorphemic [t]/[d] would entail no loss of information. An alternative explanation is formal, having to do with properties of the structural linguistic context in which the phonic variable occurs rather than with its discourse function. In a detailed study of the morphological effects on (t/d)-deletion, Guy (1991) focused on verbs that form the past not only through the addition of [t]/[d] but also through changes to their stem (e.g. *leave/left*, *keep/kept* and *tell/told*). As Table 5.8 shows, the rate of deletion in these "semiweak" or "ambiguous" verbs is intermediate to that of monomorphemic forms and past-tense verbs. To explain this effect, Guy appeals to the theory of Lexical Phonology, in which English has two levels of morphology in which phonological rules apply. As illustrated in Table 5.9, the final [t]/[d] is present in monomorphemic forms in the underlying representation but is affixed to semiweak forms at level 1 and to regular past forms at level 2. Since the (t/d)-deletion rule applies at each level, it has three opportunities to operate on monomorphemic forms, two opportunities with semiweak verbs and only one opportunity with past-tense verbs. These differences in opportunities to

Table 5.8 Rate of (t/d)-deletion according to morphological status (Guy 1991).

Morphological status:	Rate of deletion	Total N
Monomorphemic	56%	1,441
Semiweak	39%	109
Regular past	27%	600
Total:		2,150

Table 5.9 Model of (t/d)-deletion in Lexical Phonology (adapted from Guy 1991).

	Monomorphemic	**Semiweak**	**Past**
	mist	*left*	*missed*
Underlying Representation	/ mɪst / *	/ lɛf /	/ mɪs /
Level 1	/ mɪst / *	/ lɛf + t / *	/ mɪs /
Level 2	/ mɪst / *	/ lɛft / *	/ mɪs + t / *
Phonetic Realization	[mɪst]	[lɛft]	[mɪst]

* = eligible for (t/d)-deletion

apply thus explain the differences in rates according to morphological status.

Another variable that shows morphological conditioning is Spanish (s)-deletion. Since final [s] serves to mark plural on nouns, a functional hypothesis predicts less deletion with plural [s], as in *casas* "houses", than with monomorphemic [s], as in *despues* "after". Surprisingly, Poplack's (1980a) study of (s)-deletion in Puerto Rican Spanish in New York City finds the opposite result: more deletion with plural forms than with monomorphemic forms. Poplack explains the apparent anomaly of morphological conditioning in Spanish (s)-deletion by examining the syntactic context. Because Spanish marks plural redundantly on all elements of the Noun Phrase (e.g. *las*$_{\text{1st}}$ *casas*$_{\text{2nd}}$ *blancas*$_{\text{3rd}}$ "the white houses"), Poplack tests for the effect of the position of the token in the NP (1st, 2nd or 3rd) on deletion, as well as whether deleting an [s] on one constituent of the NP would lead to more deletion in the other constituents. The results are shown in Table 5.10. Note that the syntactic position of /s/ clearly correlates with deletion: initial tokens (5.10a) are least likely to be deleted, followed by those in second position (5.10b), with most deletion in third position (5.10c).

Table 5.10 Contribution of syntactic position and mark on preceding token to (s)-deletion in New York Puerto Rican Spanish (adapted from Poplack 1980a: 376)

	Position of token in NP		
	1st	2nd	3rd
Mark on preceding token:			
No preceding token (initial)	.33		
S		.39	
Ø		.55	
ØS, SS			.39
SØ			.56
ØØ			.76

(5.10) a. la$\underline{s}$ casas blancas "the white houses"
b. las casa$\underline{s}$ blancas
c. las casas blanca$\underline{s}$

Moreover, the mark on the preceding token is also important: a preceding zero increases the chances for another zero. The combination of these two effects means that deleting an NP-initial /s/ is not likely, but if it is deleted, the following tokens within the NP are also likely to be deleted. If the NP-initial /s/ is retained, the following tokens within the NP are also likely to be retained. This formal (or, rather, counter-functional) finding suggests that (s)-deletion is sensitive to syntactic structure, such that all deletion tends to apply (or fail to apply) to all constituents within an NP.

Finally, we can ask whether discourse function has a more general effect on phonic variables. While few studies consider this perspective, a recent line of research suggests that the rate of (t/d)-deletion may be affected by the frequency with which a lexical item is used in discourse. Bybee (2000) looks at the rate of (t/d)-deletion in Chicano English (Santa Ana 1991), categorizing lexical items as high- or low-frequency based on whether they occurred more than or less than 35 times per million words (using Francis and Kucera's (1982) tabulation of frequency in various corpora). As Table 5.11 shows, high-frequency words have a significantly higher rate of deletion than low-frequency words, suggesting that the more frequently a word is used, the more likely (t/d)-deletion is to apply. However, Walker's (2008) attempt to replicate these findings in a larger dataset finds no significant correlation between the frequency and deletion. As Figures 5.3a–c show, whether we measure the frequency of a lexical item according to the number of times it occurs in the token file (Figure 5.3a), in the corpus of interviews from which the tokens were extracted (Figure 5.3b), or in Francis and Kucera (1983) (the same general tabulation of frequency used by Bybee) (Figure 5.3c), the value of Spearman's rho (ρ), a measurement of linear correlation, does not achieve significance. Walker suggests that frequency does not operate blindly but interacts with lexical structure. These results do not negate the role of frequency of usage in conditioning the variation, but they suggest that this role is more complicated than is generally thought.

Table 5.11 Rate of (t/d)-deletion for Chicano English speakers (Santa Ana 1991) in high- and low-frequency words (adapted from Bybee 2000: 70).

	% Deletion	Total N
All Words		
High frequency (>35/million)	54%	1650
Low frequency (<35/million)	34%	399
	$\chi^2 = 41.67$, $df = 1$, $p < .001$	

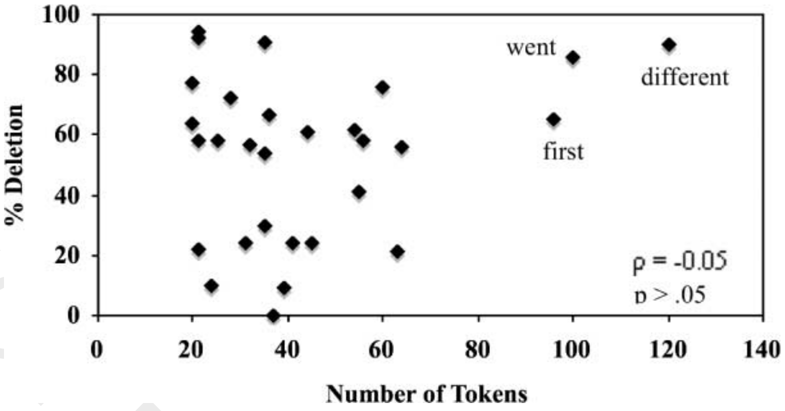


Figure 5.3a Rate of (t/d)-deletion in Toronto English, by frequency in the token file.

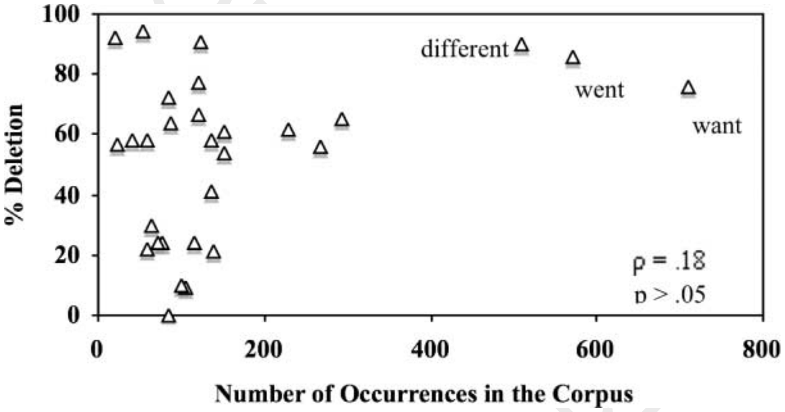


Figure 5.3b Rate of (t/d)-deletion in Toronto English, by frequency in the corpus of interviews.

5.3 Conclusion

In this chapter, we discussed the methodological and theoretical issues involved in studying variation at the level of sound systems. Rather than drawing a line between phonetic and phonological variation, we used the term phonic to refer to variation that may occur at the level of phonetics or phonology. Examining a range of phonic variables and comparing them with categorical rules and processes, we concluded that they are different in degree of application rather than in kind. Although defining the variable context for phonic variables seems relatively straightforward, it still constitutes a kind of linguistic analysis, and we need to make assumptions about the nature of the underlying forms.

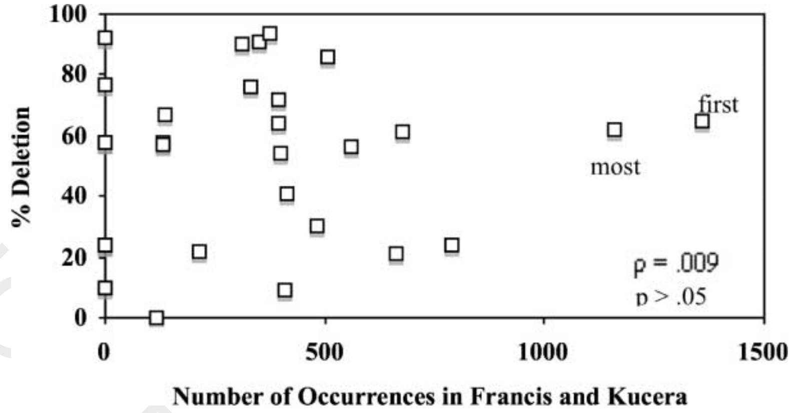


Figure 5.3c Rate of (t/d)-deletion in Toronto English, by frequency in Francis & Kucera (1982).

This consideration is important for vowel variables, whose variable context may be based on phonemes or word-classes. As continuous variables, vowels also present problems for dividing realizations into different variants and calculating relative frequencies. One approach is to make use of impressionistic or auditory analysis, another is to use instrumental or acoustic analysis, though combining both procedures is recommended.

Our focus is the language-internal conditioning of variation rather than its use in socio-symbolic functions. We discussed the different types of lexical conditioning, including individual lexical items and classes of lexical item. Although it is often difficult to draw a line between phonetic and phonological conditioning, we discussed some considerations of how this might be done, contrasting language-specific and universal conditioning. We also examined conditioning by stress as well as effects of the segmental context. We also saw evidence that phonic variables may be conditioned by grammatical factors, such as the morphological status of the phonic variable and the morphological structure and grammatical position of the word in which it occurs. We contrasted formal and functional hypotheses. Finally, we considered whether discourse function has an effect on phonic variables, looking specifically at the role of frequency.

Although phonic variation has received more attention than grammatical variation, we have seen that there are a number of methodological and analytical issues that need to be taken into consideration. These issues become more important in moving above and beyond phonology, which we do in the next chapter. As we will see, studying variation in grammatical systems will involve making some changes to some of the assumptions and methods developed in studying variation in sound systems.

5.4 Further Reading

Adank, Patti, Roel Smits & Roeland van Hout. 2004. A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America* 116(5): 3099–3107.

Boersma, Paul & David Weenink. 2008. *Praat: Doing Phonetics by Computer*. (http://www.praat.org)

Labov, William. 2008. Plotnik. (http://www.ling.upenn.edu/~wlabov)

Labov, William, Malcah Yaeger & Richard Steiner. 1972. *A Quantitative Study of Sound Change in Progress* (Vol. 1). Philadelphia: U.S. Regional Survey.

6 Variation in Grammatical Systems

6.0 Introduction

In the previous chapter, we discussed the issues surrounding phonic variables, including the definition of the variable context and the different factors that may condition the variation: lexical, phonological and grammatical. In this chapter, we move above and beyond phonetics and phonology, to variation at the level of grammatical systems. Here, and in the rest of the book, we will use the term “grammatical” to refer to linguistic systems that are normally subdivided into fields such as morphology, syntax and discourse or pragmatics. There are two reasons for conflating these fields into one level. First, many variables cannot easily be separated into different fields. Rather, they are interleaved with each other, such that variants of a variable may be distributed across morphology, syntax and discourse (and, in some cases, phonology). Second, many of these variables raise similar methodological issues that differ from those raised by phonic variation. In particular, extending the study of variation into the realm of grammar requires us to reconsider the methods used in defining the variable context for phonic variables. Not only do we need to reconsider the nature of the relation between variants, but we also need to question whether the notion of the variable rule is appropriate at this level.

We begin by discussing different types of grammatical variation, starting with grammatical variables that interact with the phonological system, before proceeding to variables that are more strictly morphological, variables that cut across morphology and syntax, purely syntactic variables and, finally, variables that cut across all of these systems but fulfill functions at the level of discourse or pragmatics. We then discuss the methodological issues that are relevant to grammatical variation, focusing on the central question of how to define the variable context. We consider the different types of factors that have been hypothesized to condition the variation: phonological factors, including the phonological and prosodic context, and grammatical factors, including morphological class, morphological function, the position and presence of other constituents, and semantic or pragmatic factors.

6.1 Types of Grammatical Variable

When we move beyond phonetics and phonology, we find variation in a number of different areas of the grammar. At first glance, some variables appear to be phonic, in that they affect a single phonic segment. For example, in all varieties of English, the realization of the verb *be* (also referred to as the copula) varies between a full syllabic form (*am/is/are*) (6.1a) and a contracted consonantal form (*'m/'s/'r*) (6.1b). Additionally, some varieties feature a third, zero variant (6.1c).

(6.1) a. The way the world is going and the way people is acting, I don't know, it don't seem like half of them believe in any God. (NS101/1A: 358)
b. The way the world's going, I don't think half of them believe in the God. (NS101: 350).
c. So, hey, this is my home, home where- this Ø where I was born. (AN8: 150)
(Walker 2000a: 55)

Since this variation involves the loss of a single phonetic segment (a vowel in the case of contraction and a consonant in the case of zero), we might be led to view this variable as phonic.

The difference between grammatical variables and truly phonic variables lies in the locus of the variation: phonic variables are defined in terms of a phonetic or phonological context, whereas grammatical variables are defined within a morphological, syntactic or discourse context. For example, the variable context for the copula is defined in terms of a grammatical context, non-verbal predication. The verb *be* occurs when the predicate is not a tensed verb. This is not a variable context, because the variation does not occur in all contexts of word-final [m], [s] and [r], but only those contexts in which those segments are realizations of the copula.

Other types of variation affecting a single segment are more clearly morphological. For example, in some varieties of English, the verbal suffix –s occurs variably in grammatical persons other than third singular (6.2). That this variation is morphological rather than phonological can be demonstrated by variation in irregular as well as regular verbs (6.3).

(6.2) If I ***go*** to a lake in- uh- in a car and a lake is handy, I ***gets*** all nerved up. (AN16: 35–6)
(6.3) a. Today, vehicles ***is*** riding on them. (BQ1: 352)
b. We not no dog we ***are*** people. (BQ1: 302)

Thus, this is not simply variation in the affixation of a single segment, but rather variation in the morphological realization of agreement in number and person between the subject and the verb.

Grammatical variables may also cut across morphology and syntax. For example, in French, events or states that will occur in the future occur variably in at least two forms: morphologically, through verb suffixation (6.4a), and morphosyntactically, through a periphrastic construction with the verb *aller* "go" (6.4b).

(6.4) a. J'ai dit, "laisse faire, on ***ir-a*** à messe demain matin".
(OH70: 686)
I-have said, "let do, one go-FUT to mass tomorrow morning"
I said, "Leave it, we'll go to mass tomorrow morning."
b. Bien demain, tu ***vas aller*** au bingo, tu ***vas gagner***.
(OH65: 2301)
well tomorrow, you go-2SG go-INF to-the bingo, you go-2SG win-INF
"Well tomorrow, you're going to go to bingo, you're going to win."
(Poplack & Turpin 1999)

Such variation, between periphrastic (morphosyntactic) and non-periphrastic (morphological) constructions, also occurs in English. For example, many contexts of present tense feature alternation between the (morphological) simple present and the (morphosyntactic) progressive (6.5 to 6.7).

(6.5) I try my best to serve my master, I'm I- trying my best to serve my uh, heavenly father, try- trying my best to serve God.
(ES11: 139–41)
(6.6) Some is living in Bronx . . . and then they- some they- they live in Manhattan. (SA4: 269–70)
(6.7) a. They're having to be fashion accessories. (TO25: 303)
b. You have to be there all the time. (TO25: 326)

Variation may be more straightforwardly syntactic, involving variable positioning within the clause or sentence. For example, clitic object pronouns in Cypriot Greek appear variably before (6.8a–b) or after (6.8c–d) the verb.

(6.8) a. *pos ton elalusan*
how him they-called
"What is his name?"
b. *tora to thimithika*
now it I-remembered
"I just remembered it."

c. . . . *oti legan tin*
. . . that they-called her
“. . . that they called her.”

d. *tora perase mu*
now it-passed me
“I’ve gotten over it now.”

(Pappas 2008)

Other forms differ not in terms of alternation between morphological or morphosyntactic constructions or in terms of position in the clause, but rather between alternate (morphological or morphosyntactic) constructions, often drawn from different periods of historical development within the language. For example, English has (at least) two periphrastic forms that are used to refer to future time, a modal construction with *will* (6.9a) and a semi-modal construction with *going to* (6.9b), each of which became available as options at different points in the history of English.

(6.9)[1] a. And he’ll probably live ’til a hundred. (QC029: 1480)
b. My doctor tells me I’m going to live ’til a hundred.
(QC029: 341)

In Brazilian Portuguese, the first person plural pronoun *nós* (6.10a), inherited from Latin, varies with a new pronoun, *a gente* (6.10b), which developed from a noun phrase originally meaning “the people”.

(6.10) a. *Bom, nós não temos condução própria.*
well, we NEG have-1PL transportation own
“Well, we don’t have our own transportation.”
b. *Então a gente depende do ônibus.*
so the people depend-3SG of-the bus
“So we depend on the bus.”

(Zilles 2005: 25)

Just as we saw with phonic variables, some grammatical variables consist of overt and null realizations of a grammatical element. We have already discussed zero copula in some varieties of English. However, grammatical variables with zero variants also exist in standard English. For example, subordinate clauses are variably introduced with (6.11a–b) or without (6.11c–d) the complementizer *that*.

(6.11) a. And I let it slip that Darth Vader was Luke’s father.
(QC71: 468)
b. She said that her father was the rector of St. Michael’s Church. (QC3: 162)
c. I can’t even believe Ø I just said that. (QC59: 1840)

d. She said Ø she used to play in- in Sillery- in the Brewar swamp. (QC3: 164)
(Torres Cacoullos & Walker 2009a)

In so-called "pro-drop" languages, such as Spanish, subject pronouns are variably realized as overt (6.12a) or null (6.12b).

(6.12) a. Ella trabajaba en una tienda.
she work-PSTPROG-3SG in a store
"She was working in a store."
b. Ø cuidaba los niños.
Ø look-after-PSTPROG-3SG the children
"(She) was looking after the children."
(Hoffman in preparation)

Finally, a number of variables cannot be neatly classified as morphological, morphosyntactic or syntactic. While such variables may involve elements of morphosyntax, rather than conveying grammatical information, such as tense, person or subordination, they fulfill functions that are more relevant to the purposes of discourse or pragmatics. For example, when quoting the speech of others, speakers of English have a number of options available to them. They may use a verb of saying, such as *say, yell, scream*, etc. (6.13a), a present or past form of the verb *go* (6.13b) and a more recent form involving the construction *be like* (6.13c).

(6.13) a. And she said, "We have to go and visit with her."
(TO27: 168)
b. And then dad he goes, "Get out just in case uh- it's uh-leaking gas or whatever and ignites." (TO16: 184)
c. And I'm like, "Okay, no wonder you're like this to me."
(TO56: 290)

Unlike phonic variables, it is less appropriate to think of this alternation as the result of a rule that derives one (underlying) form from another. This difference between phonic and grammatical variants requires that we rethink the notion of variable rules. First, however, we examine the consequences of the relation among variants to the definition of the variable context.

6.1.1 Defining the Variable Context

The study of grammatical variation presents a set of methodological and analytical challenges that differ from those of phonic variation, which variationist research has had to adapt to. First, on a more practical level, grammatical variables tend to occur less frequently in spontaneous speech than do phonic variables. A two-hour sociolinguistic interview with a single informant may yield several hundred tokens of (t/d)-deletion, whereas there may be only a few dozen tokens of reference to the future. The development of large-scale "mega-corpora" (e.g. Poplack 1989) has helped to reduce the problem of infrequent grammatical variables, but some studies have resorted to ingenious methods of data collection in order to arrive at a token file that is large enough for quantitative analysis. For example, in Harvie's (1998) study of null subject in English, which occurs extremely infrequently in English, she extracts each token of null subject and takes the subject of the clauses occurring before and after the clause with the null subject as two additional tokens. Although this results in an artificial overall rate (33%), it allows her to discover robust conditioning of the variation. In a study of pronoun variation in coordinate phrases in object position (6.14), Angermeyer and Singler (2003) not only use sociolinguistic interviews but also conduct experiments designed to elicit the variable, as well as noting occurrences of the variable in day-to-day interactions and watching unscripted television programs.

(6.14) a. There's no, like, DNA difference between you and I. (TO22: 716)
b. And then she got my other sister and me. (TO25: 819)

Undoubtedly, though, the biggest challenge in studying grammatical variation is defining the variable context. Many studies simply extend the methods developed for phonic variation to the level of grammar: that is, identify a set of forms that vary with each other. This approach, which I will refer to as "form-based", is feasible if the set of variants can be closed off. However, recall the principle of accountability, which requires that we account not only for the form that interests us, but also for all other forms with which that form varies. In some cases it may not be possible to close off the set of variants using a form-based approach. For example, in studying variation in reference to future time in English, we may take a form-based approach and note the alternate use of the modal *will* (6.9a) and periphrastic *going to* (6.9b). However, these forms are used for meanings other than the future. For example, *will* expresses not only the future but also statements of general truth (6.15).

(6.15) A boat'll sink on you, but a- . . . raft never sink.(AN32: 158)

Moreover, these forms do not exhaust the possibilities of reference to future time. Other possible ways of referring to the future include the simple present (6.16a) and the present progressive (6.15b), as well as other constructions (6.16c).

(6.16) a. No, I finish on the twenty-seventh of June. (MQ9: 405)
b. I only have it 'cause I'm going to the dentist tomorrow. (QC23: 2280)
c. And then he looks down, and he's about to hit the ball. (TO50: 696)

These facts raise a number of questions in defining the variable context for a study of the future in English. Should all of these forms be included? How many? Does (not) including any of these forms affect the results? Torres Cacoullos and Walker's (2009b) study of the future in English begins with a function-based definition (including *going to, will*, the simple present and the progressive) in which *going to* occurs at a rate of 43 percent. On the basis of quantitative pattern, they perform an analysis that pits present-tense forms (simple present and present progressive) against other forms, then excludes the present-tense forms to perform analysis of *going to* against *will*. In this latter analysis (equivalent to a form-based analysis), *going to* occurs at a rate of 51 percent. Thus, it is possible to combine function-based and form-based approaches, though the results will change depending on which approach is taken.

Several linguists (e.g. Lavandera 1978; Romaine 1981) have posed the question of whether grammatical variants really are different ways of saying "the same thing". Although the preceding section presented a number of variables from different areas of the grammatical system, we might ask whether these really constitute variables or whether each of the variant forms is distinguished by different nuances of meaning. In the 1970s, when the first studies of grammatical variation began to appear, Beatriz Lavandera (1977) and others raised concerns that the methods that had been developed to study phonic variation were being extended to grammar "without apology". One basis of objection was the (often implicit) assumption that, just as phonic variation could be viewed as the result of a variable transformational rule that applied with a certain probability, grammatical variation could also be conceptualized in terms of variable rules. However, while it may be plausible to postulate a variable rule of phonic deletion (e.g. t/d → <Ø> / __ #), it makes less (linguistic) sense to view variation in grammatical forms as the result of a rule transforming an underlying form (e.g. *will* → <*going to*> / __ [+future]).

In the case of grammatical variation, we need to move away from the variable rule as a (necessary) model of the relation between variants. In the first place, it may not be possible to determine which of the grammatical variants, if any, is the underlying form. For example, Gillian Sankoff's (1980) study of Montreal French identified variation of the complementizer *que* "that" (6.17a) with a zero complementizer (6.17b).

(6.17) a. A l'école on nous enseignait que les protestants c'est de pas bons.

"At school they taught us that the Protestants are no good."
b. Au début je pense Ø ça a été plutôt un snobisme.
"At the beginning I think Ø it was more a kind of snobbery."

In line with other studies of variables with a zero variant, we might propose a variable rule that deletes an underlying complementizer *que*:

que → <Ø> / VP [$_{\text{CP}}$ __

However, Sankoff notes that *que* also occurs variably after *quand* "when" and *comment* "how" (6.18), where standard French would not expect a complementizer.

(6.18) a. Tu sais comment qu'ça se passe.
"You know how (that) that happens."
b. Je sais pas comment Ø ça se fait.
"I don't know how it's done."

Therefore, it is unclear whether the variation is best viewed as *que*-deletion or *que*-insertion, or even whether this is a single variable or whether there are two variables involving similar variants.

The most controversial question posed by Lavandera is whether grammatical variants ever mean exactly "the same thing". Just as no two lexical items are ever entirely synonymous, it is possible that every difference in grammatical form constitutes a change in grammatical meaning. As we have already noted, many of the putative semantic nuances that distinguish different grammatical forms are often neutralized in spontaneous discourse (Sankoff 1988a). Moreover, since it is precisely the variation among grammatical forms that concerns us, defining the variable context on a formal basis may become circular.

Initial responses to this question tended to concentrate on demonstrating that grammatical variants are semantically equivalent. For example, in a study of the English passive, Weiner and Labov (1983) argue that active and passive sentences without an identifiable agent (6.19) can be considered as variants.

(6.19) a. The liquor closet got broken into.
b. They broke into the liquor closet.

They argue that since both sentences refer to the same "state of affairs", they have the same underlying set of truth conditions and thus are semantically equivalent. As subsequent studies have pointed out, though, while two forms may be truth-conditionally equivalent, there may be differences of meaning having to do with discourse or pragmatics. For example, passivization may be viewed as a pragmatic strategy of topicalizing constituents other than the syntactic subject. Thus, active and passive sentences may be referentially equivalent (that is, they refer to the same state of affairs), but not "mean" the same thing.

Another response is to relax the requirement of semantic equivalence by recognizing the neutralization of semantic nuances between grammatical forms in spontaneous, unreflecting discourse. For example, David Sankoff and Pierrette Thibault (1981) note variation in Montreal French between auxiliary *être* "to be" and *avoir* "have" in the periphrastic *passé composé* past tense with verbs that require *être* in standard French:

(6.20) a. *je suis tombé*
I am fallen
"I fell" or "I have fallen"
b. *j'ai tombé*
I-have fallen
"I fell" or "I have fallen"

The problem with defining the variable context for this variation is deciding which verbs can alternate between the two auxiliaries. There is a prescriptive set of verbs that take *être* (the infamous Dr. Mrs. Van der Tramp of French school lessons: *descendre, rentrer, mourir, rester, sortir, venir, aller, naître, devenir, entrer, retourner, tomber, revenir, arriver, monter, partir*), but membership in this set has changed over time, and in any case, as we saw above with *que*, the prescriptive rules of French are not necessarily reflected in colloquial Montreal usage. Sankoff and Thibault sidestep this issue by arguing that, in contrast with categorical differences in form according to context (strong complementarity), grammatical variables exhibit an inverse quantitative relationship across the speech community (weak complementarity). In other words, if we correlate forms of the *passé composé* with *avoir* and *être* across some social index (such as social class, education or access to the standard language), we should see a gradual increase in one form and a decrease in the other. As Figure 6.1 shows, this prediction is confirmed for *avoir/être*: as the social index of access to the standard language increases, the use of prescriptive *être* increases and the use of nonstandard *avoir* decreases.

This approach has the advantage of obviating the requirement for strict semantic equivalence, but it fails to satisfy the principle of accountability. Since we do not know exactly in which contexts the speaker has a choice between using *avoir* and *être*, we have no way of calculating relative frequencies. Sankoff and Thibault quantify their results by dividing the number of occurrences of *avoir* and *être* by the number of words in the text. Such a "normalization" procedure, which is widely used in historical and corpus linguistics, may control for the different lengths of

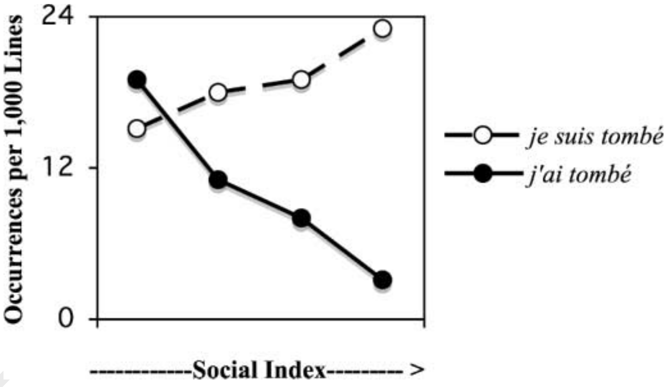


Figure 6.1 "Weak complementarity": Rates of *avoir* and *être* usage per thousand lines of transcription (Sankoff & Thibault 1981).

individual interviews, but it implicitly assumes that the variable is distributed evenly throughout discourse: that is, given any stretch of discourse, this approach predicts that the variable will always occur a particular number of times. Although to my knowledge this prediction has not been tested empirically, my experience in extracting tokens of grammatical variation from natural speech suggests that occurrences of variables tend to cluster together according to topic and discourse genre. Obviously, tokens of reference to future time will occur more frequently in a conversation about hopes for the future than in a discussion of past events. Thus, while "weak equivalence" may relax the requirement of strict semantic equivalence, it does not solve the problem of defining the variable context.

A more promising approach to defining the variable context for grammatical variation, which has been successful in more recent research, is to focus not on the semantic equivalence of variants but rather on their grammatical and discourse functions (Sankoff 1988a). This approach, which I will refer to as function-based, begins by delimiting a sector of the grammatical or discourse environment. This sector may be defined in terms of a semantic category (such as reference to past or future time or habitual aspect; e.g. Poplack & Tagliamonte 1996; Richardson 1991; Torres Cacoullos & Walker 2009b) or to a pragmatic or discourse function (such as reporting what someone else said; Tagliamonte 2006b). Note that a function-based approach necessitates re-thinking the nature of the variable context. Up until now we have defined the variable as "different ways of saying the same thing", with the variable context standing as "the same thing". In contrast with a form-based approach, in which "the same thing" corresponds to an underlying form or a set of variants, in a function-based approach, "the same thing" corresponds to a common discourse function. In this sense, the variable context can be seen less as an element of abstract structure (such as a phonological environment or a syntactic position) and more as a procedure for sifting the data. A functionally-defined variable context does not constitute a linguistic analysis in the same sense as a formally-defined variable context. Rather, the functionally-defined variable context begins by delimiting an area of research and uses the conditioning by the linguistic factor groups to further refine the analysis.

For many variables, the variable context can be approached from either a form-based or a function-based approach. We have already seen this option with the future in English, where the variable context can be defined either by enumerating a set of co-varying forms {*will, going to*} (form-based) or by examining the relative distribution of all forms expressing reference to future time (function-based). Since the decision of which approach to take depends largely on the goals of the analysis, there is no "right" or "wrong" approach. As a case in point, consider variability in the use of verbal *–s* in English. Most studies define the variable context within a form-based approach, by including as variants all bare verbs and all verbs marked with *–s* (e.g. Poplack & Tagliamonte 1991). However, many studies find an association of verbal *–s* with aspectual distinctions such as habitual activity (6.21a) (vs. continuous states (6.21b)), which are also conveyed through other, periphrastic constructions, such as the progressive and modal *will* (6.22). Moreover, since present-tense forms may be used to refer to events that occurred in the past (6.23), many unmarked verbs may in fact be past-tense verbs that have undergone another variable (phonological) process, (t/d)-deletion. A function-based approach to verbal *–s* (e.g. Walker 2000b) would include in the variable context the forms in (6.21) as well as those in (6.22), but not those in (6.23).

(6.21) a. Every time Gladys give me soup on the table it puts me right in mind. (AN32: 402)
b. Well there, the sister- she lives with her sister. (SA1: 166)

(6.22) a. If they know you handling money well then they- they raise your- your wages. (SA10: 1006)
b. People will take me for people in St. Vincent. (BQ14: 56)

(6.23) a. When she climbs up the mountain, [she was] just sliding. (SA5: 802)
b. Ashes out the stove. Stay in the house for nine days. That was way back. [. . .] You know how you burn wood? (AN14: 629)

Function-based approaches to defining the variable context for grammatical variation crucially rest on the ability to isolate a particular (discourse or grammatical) function conveyed by grammatical forms. Identifying this function remains a challenge for many variables, especially those at the level of discourse. A popular subject among students is the discourse marker *like*, which, as (6.24) shows, occurs fairly frequently in speech.

> (6.24) Just, a lot of different types of people, like, different- different races, different ages, lots of different people always in, like, every neighborhood that I lived in.
>
> (TO24: 10–12)

Does *like* constitute a variable? If so, how do we define its variable context? A form-based approach would involve enumerating all the other forms with which *like* varies, though it is unclear what these other forms are. Perhaps *like* varies with zero (i.e. realization vs. non-realization of like, as we saw with the copula), but since *like* can potentially occur in almost any syntactic position, the principle of accountability would require us to count not only every occurrence of *like*, but also every potential syntactic position in which it did not occur. One option is to "normalize" the occurrence of like per length of interview, as Sankoff and Thibault (1981) do for *avoir/être* variation. Similarly, Vincent and Sankoff's (1992) study of discourse markers in French (what they call "punctors"), such as *là* "there" and *tu sais / vous savez* "you know", shown in (6.25), normalizes the occurrence of each form per 10,000 words of transcribed interview. However, as noted above, this approach makes the (untested) assumption that the contexts in which discourse variables occur are distributed evenly throughout speech.

> (6.25) a. Franchement, là, il y en a c'est décourageant.
> "Really, there, there are so many it's discouraging."
>
> b. On savait comment vivre. On savait tu sais: c'était cette délicatesse qu'on avait, vous savez.
> "We knew how to behave. We knew, you know, it was that refinement that we had, you know."
>
> (Vincent & Sankoff 1992)

In contrast, a function-based approach to discourse variables faces the challenge of isolating the discourse function (or functions) conveyed by variants, as well as determining which other forms (or absence of forms) convey the same function(s). D'Arcy's (2005) study of *like* attempts to address these issues by taking a form-based approach, restricting the variable context to sentence-initial position (or, in her formulation, the left edge of CP), which represents an improvement over "normalization" and obviates the need to identify discourse functions (at least at the level of extraction). However, she does not count every potential context in which *like* could occur (i.e. the beginning of every sentence or CP) but does not: rather, she uses a randomly-selected subsample of sentences.

Although this approach, like Harvie's (1998) for null subject in English, allows for the conditioning by language-internal factors to emerge, it creates an artificial overall rate. Defining the variable context for discourse variables thus represents a relatively unexplored area of variationist research.

6.2 Conditioning of Grammatical Variation

In the preceding section, we considered the issues involved in defining the variable context for grammatical variation, distinguishing broadly between form-based and function-based approaches, with the goal of quantifying the variation. However, the ultimate goal of variationist analysis is not only to calculate the relative frequency with which grammatical variants occur, but also to determine their roles in the linguistic system. In categorical approaches to linguistics, any difference in form entails a difference in meaning. Thus, in this approach, determining the role of a form in the linguistic system involves determining the meaning of that form when the linguistic context changes. In contrast, the variationist approach uses the distribution of a form (that is, its probabilistic associations with different elements of the linguistic context) to infer its role in the linguistic system.

As we saw with phonic variation, grammatical variation may be conditioned by elements from different levels of the linguistic system. These levels are not always entirely distinct from each other, and usually act simultaneously in conditioning the variation, but in the following sections we will discuss each level in turn. We will begin by discussing the conditioning of grammatical variation by the lexicon, including not only individual lexical items but also frequent collocations. We then proceed to a discussion of phonological conditioning, including not only the segmental context but also suprasegmental considerations, such as prosodic structure and stress. Finally, our discussion of the grammatical conditioning of grammatical variation will divide factors into two broad classes, having to do with structural and functional considerations.

6.2.1 Lexical Conditioning

The effects of the lexicon on grammatical variation are not as well explored as they are for phonic variation. In fact, studies of grammatical variation typically control for the effects of "lexicalized" forms by excluding potential tokens that occur in fixed or "frozen" expressions. For example, studies of verbal *–s* usually exclude tokens of discourse markers such as *I mean* and *you know*, since the verbs in such expressions do not tend to vary in their morphological form.

Only recently have studies begun to explore the effects of lexical conditioning on grammatical variation. An early study in this vein is

Poplack's (1992, 1997) work on Ottawa-Hull French, in which verbs in "subjunctive-selecting" subordinate clauses vary between subjunctive (S) (6.26a) and indicative (I) (6.26b) morphology.

(6.26) a. J'espère qu'ils soient (S) pas trop ingrats . . .
(OH15: 887)
"I hope that they are not too ungrateful . . ."
b. Mais j'espère que l'Église est (I) pas contre moi pour ça.
(OH53: 1525)
"But I hope that the Church doesn't hold that against me."
c. Fallait qu'elle répond (I) "oui, tu peux faire trois pas de géant." Fallait qu'elle réponde (S) la phrase complète.
(OH25: 2186)
"She had to say 'yes, you may take three giant steps.' She had to say the whole sentence."
(Poplack 1997: 288–9)

Prior work on variation in the French subjunctive adduced different nuances of meaning between the subjunctive and indicative variants, as well as attributing this variation to contact with English, in which the subjunctive is (virtually) nonexistent. However, Poplack's results, shown in Table 6.1, reveal that most of the variation can be accounted for by the effects exerted by a small set of matrix-clause verbs that highly favor subjunctive morphology on the verb in the subordinate clause: *falloir* "have to", *vouloir* "want" and *aimer* "like". Moreover, these three verbs make up the vast majority (73%) of matrix-clause verbs in the data. Thus, not only do these verbs inflate the overall rate of the subjunctive, but their disproportionate share of the data also overshadows the effects of other matrix-clause verbs.

Table 6.1 Distribution of verbal matrices across categories of text frequency and propensity to select subjunctive mood (Poplack 1997: 293).

HIGH FREQUENCY/HIGH SUBJUNCTIVE		% SUBJUNCTIVE	% DATA
falloir	"have to"	89	62
vouloir	"want"	91	11
aimer	"like"	67	
HIGH FREQUENCY/LOW SUBJUNCTIVE			
croire (neg)	"not believe"	13	15
penser (neg)	"not think"	14	
admettre	"admit"	9	
avoir l'air	"seem"	0	
espérer	"hope"	21	
LOW FREQUENCY/VARIABLE SUBJUNCTIVE			
All other verbal matrices			12

In a similar vein, Torres Cacoullos and Walker (2009a) examine the conditioning of the zero complementizer in English (as shown in 6.27, reproduced from 2.10), which is claimed to be conditioned by the frequency and semantic class of the matrix verb. As Table 6.2 shows, in three independent variable-rule analyses of zero complementizer, lexical frequency (the third column) is selected as significant, with high-frequency matrix verbs favoring zero. The semantic class of the matrix verb (first column) is also significant, with verbs of attitude and utterance favoring zero. However, individual lexical type is also selected as significant, with *think, remember* and *say* most favorable to zero. These results suggest that the effects of frequency and semantic class may mask effects which are purely lexical. Since these factors groups are highly interrelated, we use multivariate analysis to disentangle their effects to determine which best accounts for the observed variation. Table 6.2 shows that analysis including lexical type features the log likelihood closest to zero, indicating the best fit to the data. Comparing the log likelihoods of all three analyses shows that this difference is statistically significant. Torres Cacoullos and Walker (2009a) conclude that the effects of frequency and semantic class reflect the effects of particular matrix verbs.

(6.27) a. Everyone thinks Ø I'm from Montreal. (MQ67: 1778)
b. Anybody that comes here knows that I don't speak it. (QC57: 1408)

Table 6.2 Comparison of factors contributing to zero complementizer (adapted from Torres Cacoullos & Walker 2009a).*

	Semantic Class of Matrix Verb		**Lexical Type of Matrix Verb**		**Lexical Frequency of Matrix Verb**	
Input:	.677		.691		.680	
	Attitude	.59	*think*	.77	High	.61
	Utterance	.56	*remember*	.69	Low	.36
	Knowledge	.40	*say*	.60	Medium	.35
	Suasive	.36	*know*	.43		
	Extraposition	.29	*tell*	.41		
	Comment	.28	Other	.35		
			find	.30		
			realize	.21		
Log likelihood:	−870.994		−833.587		−861.503	
		$\chi^2 = 74.814$ df = 2 p < .001		$\chi^2 = 55.832$ df = 5 p < .001		

* Other factor groups selected as significant are not shown but remain constant across analyses.

We can extend the notion of lexical effects to the phrasal level. Torres Cacoullos and Walker (2009a) also examine the conditioning of the zero complementizer by frequent collocations of matrix-clause verbs and subjects. As Table 6.3 shows, some matrix-clause subject-verb collocations, such as *I think* and *I guess*, not only constitute a large proportion of their respective lexical types (61%–99%) but also favor zero complementizer at much higher rates (76–100%) than do other subject-verb collocations (69%). In other words, a great deal of the variation in the use of the zero complementizer is conditioning by the lexicon, not only in terms of particular verbs in the matrix clause, but also by particular, frequent subject-verb collocations.

Such studies raise questions about the degree to which putative conditioning by other factors (structural or functional) can be attributed to lexical effects, whether thought of in terms of individual lexical items or in collocations of such items.

6.2.2 Phonological Conditioning

The phonological conditioning of grammatical variation has received somewhat more attention than has lexical conditioning. In fact, phonological conditioning is sometimes used as a diagnostic of the status of variables as phonic or grammatical. For example, the absence of consistent phonological constraints on verbal *–s* within and across different varieties of English has been adduced as evidence against its status as a phonic variable.

Phonological conditioning has also been used as evidence for the presence of underlying forms. Table 6.4 shows Walker's (2000b) analysis of phonological conditioning of zero copula in two diaspora varieties of African American English. The effects of the segmental phonological context are strongest, with following and preceding consonants both

Table 6.3 Subject-verb collocations by lexical type and rate of zero complementizer (adapted from Torres Cacoullos & Walker 2009a).

Frequent collocations	N	% Lexical Type	% Zero
I think	734	61	95
I guess	163	99	97
I remember	90	96	96
I find	59	66	76
I'm sure	40	74	90
I wish	17	85	100
I hope	15	79	93
Total	1118	68	92
Other collocations	216	19	69

Table 6.4 Phonological factors contributing to the occurrence of zero copula in two diaspora varieties of African American English (adapted from Walker 2000b).

	Nova Scotia		**Samaná**	
Corrected mean:	.307		.186	
Total N:	215		274	
Following Phonological Environment				
Consonant	**.58**		**.70**	
Vowel	.31		.21	
Range:		*27*		*49*
Preceding Phonological Environment				
Consonant	**.65**		**.80**	
Vowel/[r]	.44		.46	
Range:		*21*		*34*
Prosodic Structure				
Two Phonological Phrases	**.60**		**.77**	
Single Phonological Phrase	.45		.47	
Range:		*15*		*30*

strongly favoring zero. Labov (1969) used results such as this to argue that the alternation between zero copula and full or contracted copula involves deletion (rather than insertion) as a (variable) strategy for reducing phonological complexity.

The analysis reported in Table 6.4 also introduces another factor in the phonological conditioning of the copula, though above the level of phonetic segments. The prosodic conditioning of zero copula is analyzed on the basis of the Phonological Phrase (PPh), which is defined in terms of the lexical-phrasal constituents Noun Phrase, Verb Phrase and Adjective/Adverb Phrase. Depending on their position in the sentence and their interaction with adjacent elements, function words such as the copula are prosodically absorbed within the preceding or following PPh. Walker divides sentences into two types, as shown in (6.28). The crucial difference between sentences consisting of a single PPh (6.28a) and those consisting of two PPhs (6.28b–c) is the presence of a PPh boundary between the copula and the subject.

(6.28) *Single Phonological Phrase*
a. [You'*re* (going) in (debt)] (ESR6: 15)
Two Phonological Phrases
b. [(Tansy)] ['*s* (really) (good)]. (AN14: 293)
c. [The (milk) in (town)] [*is* (fifteen)]. (SA3: 69)

Although the segmental phonological context is strongest, prosodic structure is also significant, with sentences consisting of two PPhs disfavoring zero copula. Walker uses this conditioning to argue for the PPh effect as an additional reflection of reducing phonological complexity: since the PPh boundary impedes the resyllabification of the copula necessary for contraction, copula deletion is another option.

Other studies have similarly appealed to suprasegmental conditioning of grammatical variation, though usually in terms of stress. As I noted earlier, reference to first person plural in spoken Brazilian Portuguese (BP) varies between *nós* and *a gente*. Verbs with *nós* prescriptively take the first person plural ending *–mos* (6.29a) and verbs with *a gente* take the (null) third person singular ending (6.29c). However, in informal spoken BP, there is also variation in the morphology of the verb with both subjects (6.29b, 6.29d).

(6.29) a. *nós falamos*
"we speak (1PL)"
b. *nós fala*
"we speak (3SG)"
c. *a gente fala*
"the people speak (3SG)"
d. *a gente falamos*
"the people speak (1PL)"

Naro et al. (1999) propose that variable agreement is conditioned by the degree of phonological opposition between the two verb forms, depending in part on conjugational paradigms and on differences in stress and vowel quality. Table 6.5 shows their proposed hierarchy of "phonic salience" (*a salencia fônica*), which features increasing phonological opposition between the two variable morphological forms.

Table 6.6 shows their analyses of first person plural ending *–mos* for

Table 6.5 The hierarchy of phonic salience for subject-verb agreement in Brazilian Portuguese (Naro et al. 1999).

Example	Description
1 *falava/falávamos* "we spoke"	The opposition –V/-V-*mos* is unstressed in both forms
2 *fala/falamos* "we speak" *trouxe/trouxemos* "we brought"	The opposition –V/-V-*mos* is stressed in one of the forms
3 *está/estamos* "we are" *tem/temos* "we have"	The opposition –V/-V-*mos* is stressed in both forms
4 *comeu/comemos* "we ate" *partiu/partimos* "we left" *vai/vamos* "we go" *foi/fomos* "we went" or "were"	The opposition –V/-V-*mos* is stressed in both forms, and the 3rd sg form shows a diphthong with an upglide that does not appear in the plural
5 *falou/falamos* "we spoke" *é/somos* "we are"	The opposition –V/-V-*mos* is stressed in both forms, and the stressed vowel changes

four different age-groups.[2] Phonic salience is significant for all age-groups, with increasing phonic salience favoring *–mos*. These results provide support for the effect of phonic salience on variable agreement in BP. More generally, they provide further evidence that grammatical variation may be conditioned by phonological considerations, not only at the level of the segmental context, but also in terms of suprasegmental structure, such as prosody and stress.

6.2.3 Grammatical Conditioning

Most attention in studies of grammatical variation has focused on grammatical conditioning, which we may divide broadly into two types, structural and functional. Although these considerations are not easily disentangled (and some would argue that they should not be), they test rather different hypotheses that are derived from different research tradition and divergent views of the grammatical system.

Structural Conditioning: Morphology and Syntax

In this section, we provide an overview of different structural factors that have been found to condition grammatical variables. By "structural" we mean elements of the morphological and syntactic context. Structural considerations may involve the morphological class of the variable or its variants, the syntactic status of the variable or its role in the sentence, the status and role of other constituents in the sentence, and the presence or absence of other syntactic constituents in the sentence.

The first structural property we consider is the syntactic role that the variable plays in the sentence. For example, English relative pronouns vary between a *wh*-expression (*who*, *which*) (6.30a), *that* (6.30b) and a null variant (6.30c).

Table 6.6 Factors contributing to the occurrence of first person plural agreement in first person plural contexts in Brazilian Portuguese, in four age-groups (Naro et al. 1999).

	Older Adults	**Younger Adults**	**Young People**	**Children**
Phonic salience				
2	.21	.07	.14	.13
3	.37	.52	.42	.60
4–5	.87	.93	.90	.82
Range:			*76*	*69*
Tense				
Present	[ ]	[ ]	.25	.13
Preterit	[ ]	[ ]	.75	.87
Range:			*50*	*74*

(6.30) a. Because I had a lot of family who lived in the close area. (TO25: 27)
b. I had to deal with things that I never thought- never even thought about. (TO27: 157)
c. Well, there's a lot of things Ø I can eliminate, but . . . (TO13: 597)

Many studies of zero relativizer have shown that the syntactic role of the relative pronoun in the matrix clause is a major consideration, with a major division between subject and non-subject relatives. This finding is confirmed in Table 6.7, which is adapted from Levey's (2006) study of relativization in adolescent London English. As Table 6.7 shows, syntactic role is the most important factor group, with subject relatives highly disfavoring the zero (with a factor weight of .27).

Another structural consideration is the syntactic role and status of other constituents in the sentence. Marking of morphological agreement on the verb is usually considered to be achieved on the basis of relations between the verb and its (syntactic or semantic) subject. Thus, variable agreement is often conditioned in part by the type of subject, either in terms of its status (pronoun or NP), its grammatical person and number, or its proximity to the verb. Table 6.8 presents results from Van Herk and Walker's (2005) analysis of verbal *–s*-marking in letters written by semiliterate African Americans who had emigrated to Liberia. For both irregular (*be, have, do*) and regular verbs, the grammatical person, subject type and adjacency of the verb all contribute to *–s*-marking, which is favored in third person singular and plural and with subjects other than adjacent pronouns.

For other variables, not only the subject is important but also the grammatical constituent following it. The copula is a variable for which both considerations are important. Table 6.9 shows an analysis of zero

Table 6.7 Factors contributing to the occurrence of the zero relativizer in adolescent London English (adapted from Levey 2006) (N=183).

		%	N
Syntactic role of relative marker			
Subject	.27	8	111
Object	.81	54	61
Oblique	.87	64	11
Range:	*60*		
Length of relative clause			
Less than 5 words	.65	34	141
More than 5 words	.12	2	42
Range:	*53*		

Table 6.8 Factors contributing to the occurrence of –*s*-marking in semiliterate African American letters from Liberia, by verb type (writers from non-Deep South States only) (Van Herk & Walker 2005).

		Regular Verbs	Irregular Verbs
	Total N:	417	630
	Input:	.126	.282
Grammatical Person			
3rd sg.		**.88**	**.91**
3rd pl.		**.70**	**.52**
Non-3rd		.08	.12
	Range:	*82*	*79*
Subject Type + Adjacency			
Adjacent Pronoun		.41	.27
Other		**.67**	**.69**
	Range:	*26*	*42*

Factor groups not selected as significant: Aspect.

Table 6.9 Factors contributing to zero copula in two different communities on Bequia, St. Vincent and the Grenadines (adapted from Meyerhoff & Walker 2007).

		Hamilton	Mt. Pleasant
	Total N:	1002	640
	Input:	.386	.459
Following grammatical category			
gonna		.90	.83
Verb-*ing*		.82	.79
Adjective		.64	.47
Prepositional Phrase		.38	.53
Noun Phrase		.16	.12
Locative Adverb		.08	.53
	Range:	*82*	*71*
Subject type + preceding segment			
NP, Vowel		.58	.87
Pronoun, Vowel		.53	.47
NP, Consonant		.43	.55
	Range:	*15*	*40*

copula in two communities on the island of Bequia (St. Vincent and the Grenadines) (Meyerhoff & Walker 2007). In both communities, both the type of subject and the following grammatical category are significant, though the relative effects for each factor reveal interesting differences between communities. In the following grammatical category, Mt. Pleasant shows a split between auxiliary *be* (that is, future *gon(na)* and present participle V-*ing*) (6.31a–b), which favor zero, and copular *be* (that is, predicate adjectives, prepositional phrases, noun phrases and locative adverbs) (6.31c–f), which disfavor. In contrast, Hamilton treats predicate adjectives (6.31c) more like verbal predicates, in that they favor zero copula. In terms of the type of subject, Mt. Pleasant shows a syntactic split, with NP subjects (6.31a,c) favoring zero and pronoun subjects (6.31b) disfavoring. In contrast, Hamilton shows more of a phonological effect, with vowel-final subjects (6.32a–b) favoring zero copula and consonant-final subjects (6.32c) disfavoring. This difference in conditioning between the two communities demonstrates that even varieties of language that share the variable presence or absence of a feature may differ in terms of its underlying conditioning. More generally, variables may be conditioned by several elements of the syntactic structural context simultaneously.

(6.31) a. They Ø telling you pull up. (BQ.P2: 599)
b. Going other places, Rasta Ø clean. (BQ.H5: 698)
c. The way the work is there now. (BQ.M303: 486)
(6.32) a. Yeah, I think my boy Ø gon done this year. (BQ.H5: 420)
b. He Ø making speed, running. (BQ.H3: 217)
c. They does go walk and bawl, days before they they're dead. (BQ.P14: 274)
d. So they figure everybody is for theyself. (BQ.M303: 634)
e. He Ø there in Antigua. (BQ.P19: 731)
f. But her father is a Ollivierre. (BQ.P24: 172)

In many cases, it is not the type of preceding or following syntactic constituent, but rather the presence or absence of constituents, that contributes to the conditioning of the variation. For example, the modification of the VP by adverbials often conditions variation between different morphosyntactic variants. Table 6.10 shows the effect of adverbials on two variants of future temporal reference in Torres Cacoullos and Walker's (2009b) study of Canadian English.

Table 6.10 Contribution of temporal adverbials to two variants of future temporal reference in Canadian English (adapted from Torres Cacoullos and Walker 2009b).*

	Futurate Present (vs. *will* and *going to*)	***will*** (vs. *going to*)
Temporal Adverbial		
Specific/definite	.71	.48
No adverbial	.48	.48
Nonspecific/indefinite	.42	.67

* Other factor groups included in the analysis not shown.

(6.33) a. Now tomorrow I'm going out with the girls from the
railway for lunch. (MQ9: 1147)
b. Well I- someday I'll be leaving. (QC51: 726)

Neither variant is particularly favored or disfavored by the absence of an adverbial. Futurate present is favored by a specific or definite adverbial (6.33a), while *will* is favored by a nonspecific or indefinite adverbial (6.33b). We may interpret the adverbial effect as a consequence of the range of meanings that each of the variants can enter into. In this case, both the present and *will* convey meanings other than the future. For example, the present tense refers to present time but it also may refer to timeless situations, and *will* also refers to habitual or timeless situations. Thus, the presence of adverbial modification may serve to disambiguate the futurate readings of these forms in this context.

Functional Conditioning: Semantics and Discourse

In the previous section, we presented different elements of the linguistic structure (morphological and syntactic) that have been found to condition grammatical variation. In this section, we consider the effects of functional considerations, using the term "functional" to refer to semantic or pragmatic-discourse considerations.

For many variables involving verbal morphology, semantic considerations such as tense-aspect and negation have been found to be important. In the study of variable subject-verb agreement in English, conditioning by aspectual distinctions has received a great deal of attention. Table 6.11 shows an analysis of factors contributing to verbal *–s* in two diaspora varieties of African American English (adapted from Walker 2000b). Although grammatical person exerts a stronger effect, aspect is consistently selected as significant in both varieties, with habitual contexts favoring and durative contexts either disfavoring (.44 in Samaná) or having no effect (.51 in Nova Scotia). Thus, what has historically been regarded as a marker of subject-verb agreement in these varieties has also taken on the function of conveying habituality.

Negation is another aspect of the discourse-pragmatic context that has been found to condition grammatical variation. For example, in Poplack and Turpin's (1999) study of future temporal reference in Canadian French, a major finding is the effect exerted by negation. As Table 6.12 shows, negation favors the inflected future very highly (.99).

Tracking of referents in discourse is another function that may affect grammatical variation. We have already discussed variation in the realization of subjects in Spanish as null or overt (as in example 6.12). A functional hypothesis would predict that any change in the referent of the subject from one sentence to the next ("switch reference"), as in (6.34a), would be more likely to promote an overt realization of the next subject

Table 6.11 Factors contributing to verbal *–s* in two diaspora varieties of African American English (adapted from Walker 2000b).

		Samaná	**Nova Scotia**
	Total N:	2,520	2,649
	Input:	.200	.047
Grammatical Person			
3rd singular		.73	.93
3rd plural		.56	.62
Non-3rd		.36	.32
	Range:	*37*	*61*
Aspect			
Habitual		.60	.64
Durative		.44	.51
Punctual		.27	.31
	Range:	*33*	*33*
Subject-verb adjacency			
Adjacent		[ ]	.51
Nonadjacent		[ ]	.32
	Range:		*19*

Table 6.12 Factors contributing to variants of future temporal reference in Ottawa-Hull French (adapted from Poplack & Turpin 1999).

		Inflected Future	**Periphrastic Future**
	Input:	.145	.727
	Total N:	725	2,627
Negation			
Negative		.99	.01
Affirmative		.36	.65
	Range:	*63*	*64*
Adverbial Specification			
Non-specific		.85	.19
No adverbial		.47	.56
Specific		.37	.23
	Range:	*45*	*37*
Grammatical Person			
Formal *vous*		.81	.22
Other		.49	.51
	Range:	*28*	*29*
Contingency			
Contingent		.51	
Assumed		.45	
	Range:	*6*	
Temporal Distance			
Proximal		.52	.56
Distal		.48	.43
	Range:	*4*	*13*

than when the referent does not change (6.34b), since the change of referent would carry important discourse information (Cameron 1993). This hypothesis is tested in Table 6.13, which shows the results of a study of El Salvadoran Spanish in Toronto (Hoffman, in preparation). Although grammatical person is the most important effect, a switch in subject reference disfavors null subject (or rather, favors overt subject). This result provides support for the effect of discourse factors.

(6.34) a. Anoche el teléfono que usted me dió? . . . De esa casa nosotros la llamamos.
"Last night the telephone number that you gave me? . . . From that house we called it."
b. Y entonces pero ella era tan gritona, que cuando ella lo decía, Ø lo decía tan y tan fuerte.
"And then but she was so loud, that when she would say it, (she) would say it so loud."
(Cameron 1993: 314–15)

As a final consideration, there may be discourse factors conditioning the variation that stem from the historical development of the variation. Although we will consider the process of grammaticalization in more detail in the next chapter, here we consider two examples of discourse-functional conditioning of grammatical variation that involve forms undergoing grammaticalization. Recall that first person plural in Brazilian Portuguese is variably expressed using the historical first person

Table 6.13 Factors contributing to null subject in El Salvadoran Spanish in Toronto (adapted from Hoffman, in preparation).

	Total N:	2,025	
	Input:	.67	
Grammatical Person and Number			
1st plural		.72	
3rd plural		.54	
1st singular		.48	
2nd singular		.41	
3rd singular		.41	
	Range:		*31*
Switch Reference			
Same		.60	
Switch		.38	
	Range:		*22*
Preceding Token			
Null		.55	
Overt		.39	
	Range:		*16*

pronoun *nós* or a newer pronoun *a gente* derived from a noun phrase meaning "the people". Zilles (2005) hypothesizes that the generic reading inherent in the historical lexical source of *a gente* should persist in its conditioning. As Table 6.14 shows, this hypothesis receives support: the type of reference is selected as significant, with generics favoring *a gente* (.63) over referentials.

As a final example, let us consider the quotative *be like*. This form is presumed to have originated in vocalization of internal thought in narratives (a form of evaluation of the actions in the narrative). Table 6.15 shows Tagliamonte and Hudson's (1999) analysis of *be like* quotatives in U.K. and Canadian English. In both communities, *be like* is favored by internal dialogue (6.35a) or non-lexicalized sound (6.35b) over reported speech. These results suggest that elements of the discourse context

Table 6.14 Factors contributing to *a gente* in first person plural contexts in Brazilian Portuguese in Porto Alegre (adapted from Zilles 2005).

	Total N:	1944			
	Input:	.85			
				%	N
Subject in previous clause					
a gente		.88		97	432
Null + unmarked verb		.33		73	105
Null + 1st pl. verb		.05		23	52
nós		.02		9	161
	Range:		*86*		
Reference					
Generic		.63		77	931
Referential		.38		61	1013
	Range:		*15*		
Subject-verb proximity					
SXV		.58		75	330
SV		.48		68	1591
	Range:		*10*		

Table 6.15 Contribution of content of quote to the occurrence of quotative *be like* in U.K. English and Canadian English (adapted from Tagliamonte & Hudson 1999).

	U.K.	**Canada**
Content of the Quote		
Direct Speech	.45	.47
Internal Dialogue	.57	.69
Non-lexicalized Sound	.67	.64

conditioning the variation may stem from the historical development of the variants from their lexical sources.

(6.35) a. She's like, "Right, you know, we're taking you out." I
was like, "Ah, I don't want to go out. Please, no."(UK/j)
b. And it was like, "Whoosh." (UK/K)
(Tagliamonte & Hudson 1999)

6.3 Conclusion

In this chapter, we moved above and beyond phonetics and phonology, to variation and its conditioning at the level of grammar, defining grammar widely to include not only morphology and syntax but also discourse and pragmatics. We discussed different types of grammatical variation, including interaction with the phonological system, more strictly morphological variation, variation cutting across morphology and syntax, purely syntactic variation and variation at the level of discourse or pragmatics. We distinguished between "form-based" and "function-based" approaches to defining the variable context, examining the criteria and consequences of choosing between these approaches. The difference between phonic and grammatical variation required moving away from the variable rule as a model of variation. We addressed the controversial question of whether grammatical variants are ever entirely equivalent, considering approaches that focus on semantic equivalence and weak complementarity, and a more recent approach that delimits a sector of the grammatical or pragmatic-discourse environment.

We considered the different types of factors hypothesized to condition grammatical variation. Lexical effects on grammatical variation have not been well studied, but there is evidence of conditioning by particular lexical items as well as by lexical classes and frequent collocations of lexical items. Phonological conditioning has been used as a diagnostic of the grammatical status of variables and for the presence of underlying forms. Grammatical conditioning may be divided broadly into two groups: structural and functional. Structural considerations include the syntactic role of the variable other constituents in the sentence, the type of preceding or following syntactic constituent and the presence or absence of constituents. Functional considerations refer to semantic or pragmatic-discourse considerations, such as tense-aspect and negation and elements of the discourse context stemming from historical development of the variation. We have seen evidence that all of these factors may operate independently and together in conditioning grammatical variation.

In this chapter, and the preceding four chapters, we have established the principles of variationist analysis. Especially important in variationist analysis are defining the variable context, which determines the calculation of relative frequencies, and the conditioning of variation by language-internal factors. Since this conditioning can be taken as evidence of the underlying linguistic system, we may make use of this relationship between variable conditioning and the linguistic system. In the following chapters, we will make use of the principles of variationist analysis to address issues in linguistic analysis in which both variation and membership in different linguistic systems are crucial considerations.

6.4 Further Reading

Lavandera, Beatriz. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7(2): 171–82.

Sankoff, David & Pierrette Thibault. 1981. Weak complementarity: Tense and aspect in Montreal French. In *Syntactic Change*, ed. by B. Strong Johns & D. Strong. Ann Arbor: University of Michigan, 206–16.

Weiner, E. Judith & William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19: 29–58.

7 Language Change

7.0 Introduction

The previous chapters established the principles of variationist analysis and variation at different levels of the linguistic system. Especially important among these principles are the definition of the variable context, since this determines how we calculate relative frequency of occurrence. However, we are not only concerned with overall relative frequencies but also with the conditioning of the variation by language-internal factors. We can use the hierarchy of linguistic conditioning to infer the variable linguistic system.

In this chapter, we apply the principles of variationist analysis to the study of language change. Since variation and change are intertwined, the variationist approach has advantages over other, strictly categorical models of language. Rather than viewing variation as a problem to be overcome, the variationist approach recognizes the inherent variability of language. This approach can accommodate the changes in frequency that are necessarily involved in language change. Moreover, the use of linguistic conditioning to infer the linguistic system means that changes in conditioning reflect changes in the linguistic system. We can thus use the quantitative patterning of variation to test models of language change.

Several questions recur over the course of this chapter. Is language change gradual or abrupt? If it is abrupt, how do we explain the incremental nature of change? Does language change apply across the linguistic system without exception, or is it linguistically conditioned? If it is linguistically conditioned, do differences in conditioning across time periods indicate different initial conditions or different rates of change? Is the history of a linguistic change reflected in current linguistic conditioning? To answer these questions, we consider two main approaches to the study of language change in the variationist approach. One approach, Kroch's (1989) Constant Rate Hypothesis, concerns itself with the question of whether language change proceeds exceptionlessly or whether it is constrained by elements of the linguistic context. Another approach is concerned with grammaticalization (Hopper & Traugott 2003), the process by which linguistic forms are adapted to serve new functions. We will conclude with some consideration about whether these two approaches can be reconciled.

7.1 Variation and Change

Whatever the theoretical orientation, all models of language must acknowledge that change proceeds gradually. In a change from form *x* to form *y*, there is always a period in which *x* and *y* co-exist: that is, there is always a period of variation, which any model of language change must be able to account for. The gradualness of language change presents a problem for approaches to the study of language that do not recognize variation. These approaches generally view language change as proceeding by reanalysis across generations. As illustrated in Figure 7.1, the output (i.e. the language production) of the linguistic system L_i of Generation *x* serves as the input to the linguistic system of Generation *x+1*. If the next generation converges on the same analysis (L_i), normal language transmission occurs and the language does not change. The language changes when Generation *x+1* arrives at a different analysis (L_j) of the output of the linguistic system (L_i) of Generation *x* (see, e.g. Lightfoot 1979, 1991).

If this model of language change is correct, we should expect to see change occurring suddenly, across a single generation. Yet an examination of historical data shows that linguistic changes take several generations, or even centuries, to occur. Attempts to resolve this dilemma have argued that the apparent gradualness and variation of language change represents either a set of discrete changes in subsequent generations or the co-existence of discrete, categorical linguistic systems in the same speech community. However, just as we saw in Chapter 2, variation persists no matter how thinly we slice the data. Language change advances via variation.

In the variationist approach, the recognition of the inherent variability of language allows us to incorporate variation, and therefore change, into

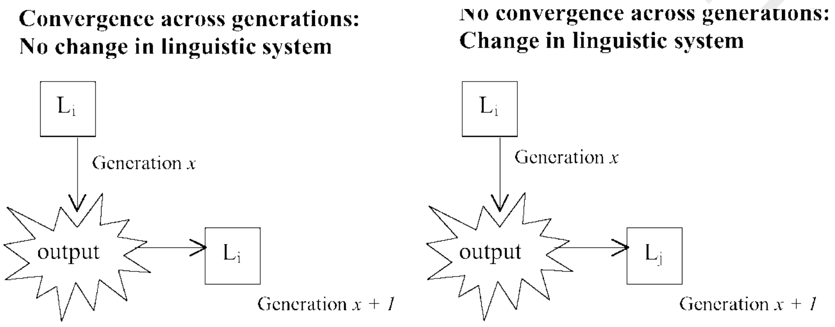

Convergence across generations:
No change in linguistic system
L$_i$
Generation *x*
output
L$_i$
Generation *x + 1*
No convergence across generations:
Change in linguistic system
L$_i$
Generation *x*
output
L$_j$
Generation *x + 1*

Figure 7.1 Scenarios of language change across generations.

the study of language. However, while all change requires variation, not all variation necessarily leads to change. Some variables exhibit stable variation, which may persist over long periods of time. For example, the variable (ing) in English can be traced to changes that took place between Old and Middle English (Labov 1989). How do we decide whether a particular variable is stable or a change in progress?

Many studies have shown that social and linguistic factors serve to initiate and impel language change, and that such conditioning can be used as evidence for or against the existence of language change. In a seminal paper written in 1968 with the late Uriel Weinreich and Marvin Herzog, Labov proposes five problems that any empirically-based account of language change needs to solve:

I. Constraints: What changes are possible? What are the conditions on change?
II. Transition: How exactly does form *x* change into form *y*?
III. Embedding: What is the a) linguistic and b) social context in which the change occurs?
IV. Evaluation: What are the subjective correlates of structural changes? How do people in the speech community view the change?
V. Actuation: How does the change begin?

Without downplaying the importance of the social aspects of language change covered by the Embedding (IIIb) and Evaluation (IV) problems, in this chapter we focus primarily on the linguistic problems of Transition (II) and Embedding (IIIa). Since, as we have seen in previous chapters, the hierarchy of linguistic constraints constitutes the variable linguistic system, we may use a change in the linguistic conditioning to infer a change in the language. For this reason, in studying language change from a variationist perspective, we are concerned not only with changes in frequency, but also with changes in the language-internal constraints. As in previous chapters, we rely on quantitative techniques of modeling, in this case to elucidate the linguistic conditioning of change.

7.2 Apparent Time and Real Time

Ideally, any study of language change involves making observations of the same language at different points in time: that is, we observe a particular language in 2009, and observe the same language again in 2019, 2029, 2039, and so on, or we go back in time and observe the same language again in 1999, 1989, 1979, and so on. Such real time studies are not always feasible, for a number of reasons. Practicality is one consideration: the timespan required for such studies is simply not realistic for most researchers. Studies of language change that has already taken place are limited by the relative recency of reliable sound-recording technology. Even where we do have recorded speech data, it usually does not represent the full range of social and linguistic contexts. To study language change that occurred before the invention of sound-recording technology, we must rely on written texts. However, only a small percentage of texts survives from any historical period, and since written language usually constitutes a genre distinct from spoken language, such texts typically do not represent the full range of speaker behavior. In addition, since until recently literacy was the preserve of a small minority of speakers, the authors of these texts do not represent the full social diversity of their communities.

For these reasons, it is more common in sociolinguistics to study language change synchronically, making use of the construct of apparent time to infer change. By studying a language at one point in time and examining the distribution of variation by age group, we may interpret differences between age groups as evidence of temporal sequencing in the past. Consider Clermont and Cedergren's (1979) study of the spread of velar /r/ (vs. apical /r/) in Montreal French, shown in Figure 7.2. Velar /r/ occurs at very low rates for older speakers (those born between 1900 and 1919), but it is the preferred variant for the youngest speakers (born between 1950 and 1959). Thus, we could interpret the distribution of velar /r/ by age group in the 1970s as a reflection of the spread of this feature in Montreal French throughout the twentieth century.

The apparent-time approach overcomes some of the practical limitations of real-time studies, but it relies on a number of assumptions. Foremost among these is the assumption that a speaker's language does not change substantially across the course of their lifespan. In order to infer language change from the distribution shown in Figure 7.2, for example,

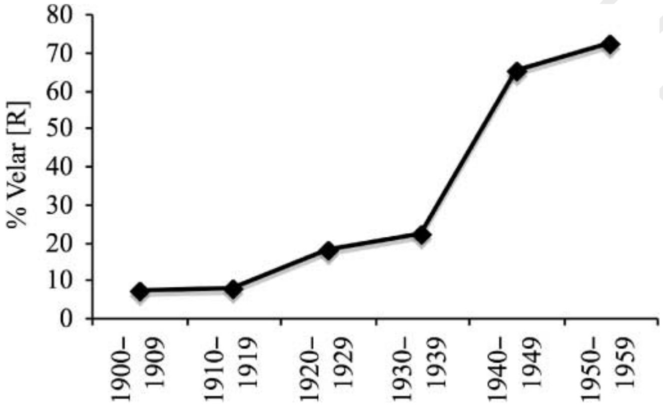


Figure 7.2 Distribution of velar [R] in Montreal French, by age group (Clermont & Cedergren 1979).

we must assume that a speaker born at the beginning of the twentieth century has not altered their way of speaking in the subsequent 70 years. There are two alternative interpretations of the distribution in Figure 7.2. One is age-grading, in which speakers change their linguistic behavior throughout their life, according to life-stages (i.e. childhood, adolescence, adulthood, old age). Under this view, if we returned to Montreal 20 years after the Clermont and Cedergren study, we would expect the people born in the 1920s to have changed their behavior to resemble that of the oldest group in Figure 7.2. Another interpretation is communal change (Labov 1994), in which speakers increase the frequency of an incoming form across their lifespan, even after childhood, in concert with an increase in frequency across the community. Sankoff and Blondeau's (2007) analysis of /r/ in Montreal provides support for the validity of the apparent-time construct. However, the possibility of age-grading and communal change present problems for the apparent-time hypothesis and the synchronic study of language change. Although we may use evidence of the linguistic and social conditioning of the change to argue for one interpretation or another, the only decisive way to resolve these problems is through a study in real time.

7.3 Testing Models of Language Change

As we saw in previous chapters, the variationist approach to the study of language makes use of quantitative modeling of linguistic data, fitting the observed data to mathematical models that represent hypotheses that we wish to test, using statistical techniques such as multiple regression and tests of significance. The statistical techniques we use depend on the mathematical models, which in turn rest on the hypotheses.

Several models of language change have been proposed, but these models can essentially be reduced to two questions:

1. Is language change sudden or gradual? This question may seem odd, given the fact of gradualness discussed above. However, it is possible that the apparent gradualness masks more abrupt changes (i.e. re-analysis within a single generation) that take time to propagate through the language or community.
2. Does language change spread exceptionlessly or is it linguistically conditioned? Studies have noted the tendency for incoming forms to occur with different frequencies in different linguistic contexts. Given the linguistic Embedding problem discussed above, we need to understand the relation between the incoming form and the linguistic context.

7.3.1 The Wave Model and the Constant Rate Hypothesis

One of the first models of language variation and change to be proposed is Charles-James Bailey's (1973) "wave model". As the first principle of this model, he notes that language change tends to take the pattern of an S-shaped curve. As shown (in an idealized form) in Figure 7.3, changes begin slowly, accelerate in their middle stages, and end slowly.

The second principle of Bailey's model (which gives it its "wave" name) states that language change proceeds in waves across (social and) linguistic contexts. Thus, differences in frequencies across linguistic contexts reflect the different times at which the incoming form spread to each context, as well as different rates of acceptance of the form in each context. As time increases, the incoming form continues to occur more frequently in early or more favorable contexts.

As Kroch (1989) points out, Bailey never tested his wave model with empirical data. In addition, the wave model does not exhaust the logical possibilities of how language change begins and spreads. Change may begin sequentially (Bailey's second principle), with the incoming form appearing first in the most favorable context and then spreading to other contexts, but change may also begin in all contexts simultaneously. If the change begins simultaneously, it may begin equally in all contexts, such that at the point of actuation there is no difference in frequencies among contexts, or it may begin differentially, such that the initial frequency differs for each context. Once the change has begun, there may be different rates of change in each context (also part of Bailey's second principle), and contexts in which the form spreads faster eventually favor the

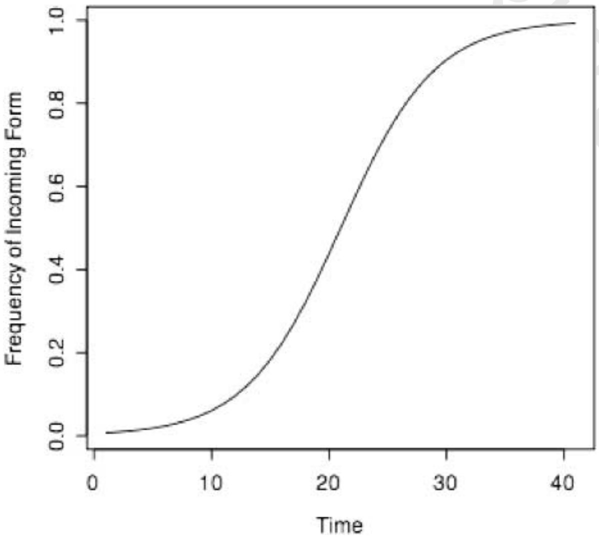


Figure 7.3 S-shaped curve of language change (idealized).

incoming form more than other contexts, or (*per* Kroch but *contra* Bailey) the change may spread at a constant rate in all contexts. According to this constant rate hypothesis, any differences in frequencies between contexts simply reflect differences in frequencies at the point of actuation of the change. How do we test these models of language change?

Determining the rate of change requires estimating the parameters of the mathematical model of the change, and testing those parameters against the predictions made by each of the above scenarios. Kroch uses the mathematical function in (7.1) to express the S-shaped curve of language change.

(7.1) *Formula for the S-shaped curve*

$$p = \frac{e^{k + st}}{1 + e^{k + st}}$$

In the function in (7.1), *p* represents the frequency of the incoming form at a particular point in time *t* (*e* is a mathematical constant, and *s* and *k* represent values that we will explore below). While the function in (7.1) plots the frequency of the incoming form, it does not tell us the rate at which the change proceeds. To determine the rate of change, we need a linear value of correlation between change in frequency (*p*) and time (*t*). A logistic transformation converts the function above into the linear function in (7.2).

(7.2) *Logistic transformation of the S-shaped curve*:

$$\ln \frac{p}{1 - p} = k + st$$

In the function in (7.2), *ln* is a "natural logarithm" (i.e. a logarithm of base *e*), *t* is time and *p* is the frequency of the incoming form at a particular point in time. The value *k* represents the intercept, the point at which the line crosses the *y*-axis at *t*=0. In other words, *k* represents the frequency of the incoming form at the beginning of the change. Of greater interest to us is the value *s*, which represents the slope of the function, the rate at which *p* changes over time: in other words, the rate of change. Thus, we can interpret any significant change in the value of *s* across different time periods as representing different rates of change. In terms of linguistic conditioning, different rates of change are also reflected in different weightings of factors across time periods. Conversely, if the rate of change is constant, *s* is similar across time periods, and the weighting of linguistic factors remains the same.

To support the constant rate hypothesis, Kroch uses a number of historical examples, a few of which we will consider. The first is the rise of the definite article with possessive NPs in Continental Portuguese (7.3), which Oliveira e Silva (1982) tracks between the fifteenth and twentieth centuries.

(7.3) a. Maria conhece meu irmão.
"Mary knows my brother."
b. Maria conhece o meu irmão.
"Mary knows (the) my brother."

As Figure 7.4 shows, the frequencies across the centuries (the white diamonds) follow the classic S-shaped curve of language change, but Kroch's logistic transformation of the same data (the black diamonds) reveals that the slope of the curve remains steady across time periods, thus providing support for the constant rate hypothesis. Oliveira e Silva (1982) also examines the effect on the definite article of four grammatical factors: whether the possessed NP is a kinship term; whether the possessive NP has a unique referent in the discourse context; whether the possessive NP is the object of a preposition; and whether the possessive pronoun is third person. As Figure 7.5 shows, the factor weights for each of these factors remains roughly the same across each time period (solid lines). When Kroch fits a regression line to each of these factors (dashed lines), he finds no significant change in effect across time periods. Thus, this case of language change provides support for the constant rate hypothesis, not only in terms of similar slopes across time-periods but also in terms of similar linguistic conditioning.

Another example discussed in Kroch (1989) is Noble's (1985) study of *have got* in British English, illustrated in (7.4b), which began to vary with *have* at very low rates before the eighteenth century, but by the twentieth century was the preferred variant.

(7.4) a. Anyhow, she has what amounts to a high Cambridge degree. (1898)
b. You've got plenty of hair. (1968)

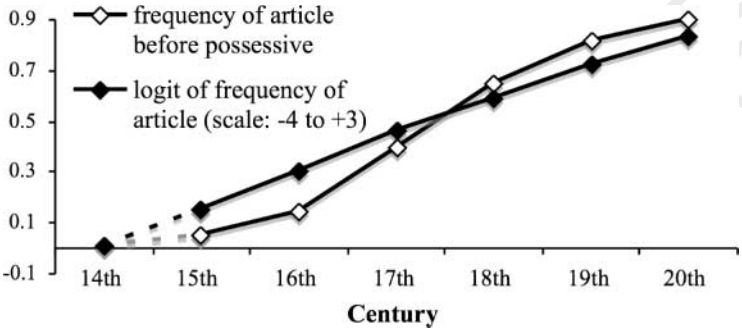


Figure 7.4 Rise in the use of the definite article before possessives in Portuguese (adapted from Kroch 1989 and Oliveira e Silva 1982).

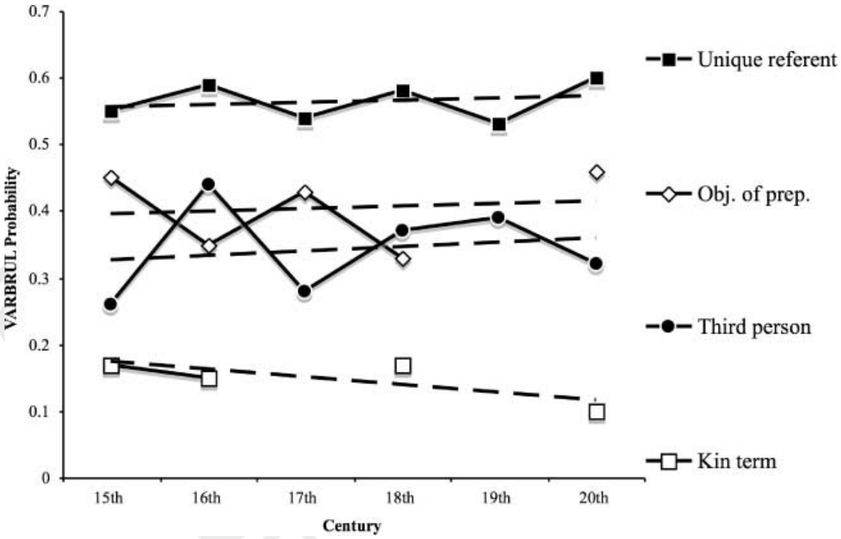


Figure 7.5 Stability of factor weights over time in the use of the definite article with possessives (adapted from Kroch 1989 and Oliveira e Silva 1982).

Noble examines the effects of two factor groups: type of possession (bounded or permanent) and type of object (concrete or abstract). As Table 7.1 shows, although the frequency of *have got* rises substantially across the three time periods, the weighting of the favoring factors changes very little: .64 to .66 for bounded possession and .58 to .66 for concrete objects. Thus, the linguistic conditioning of this change across time periods provides further support for the constant rate hypothesis.

As a final example, we examine the development of *do*-support in early Modern English, based on Kroch's (1989) reanalysis of Ellegård's (1953)

Table 7.1 Contribution of two factor groups to *have got* (vs. *have*) in British English in three time periods (adapted from Noble 1985, cited in Kroch 1989).

	1750–1849			1850–1899			1900–1935		
		%	N		%	N		%	N
Type of possession									
Bounded	.66	12	83	.64	34	99	.66	89	74
Permanent	.34	4	108	.36	16	122	.34	70	43
Range:	*32*			*28*			*32*		
Type of object									
Concrete	.66	13	68	.61	34	74	.58	86	51
Abstract	.34	4	123	.39	20	147	.42	79	66
Range:	*32*			*22*			*16*		

study. In contrast with Modern English, main verbs in Middle English could be fronted in questions and with *not* (7.5).

(7.5) a. How great and greuous tribulations suffered the Holy Appostyls?
b. . . . spoile him of his riches by sondrie fraudes, whiche he perceiueth not.

In generative grammar (or at least in the contemporaneous version adopted by Kroch), fronting is viewed as movement of the verb from its underlying position in the Verb Phrase to a higher, functional node (Infl(ection)) to receive tense-marking (Verb-to-Infl movement), as illustrated in Figure 7.6. In early Modern English, the rule of Verb-to-Infl movement began to be lost for all verbs except auxiliaries, and a dummy verb *do* began to appear in Infl in order for tense features to be realized if the main verb was not moved. Sentences such as those in (7.5) began to vary with those such as in (7.6), not only in cases where Verb-to-Infl movement was blocked by intervening elements (7.6a–b) but also where

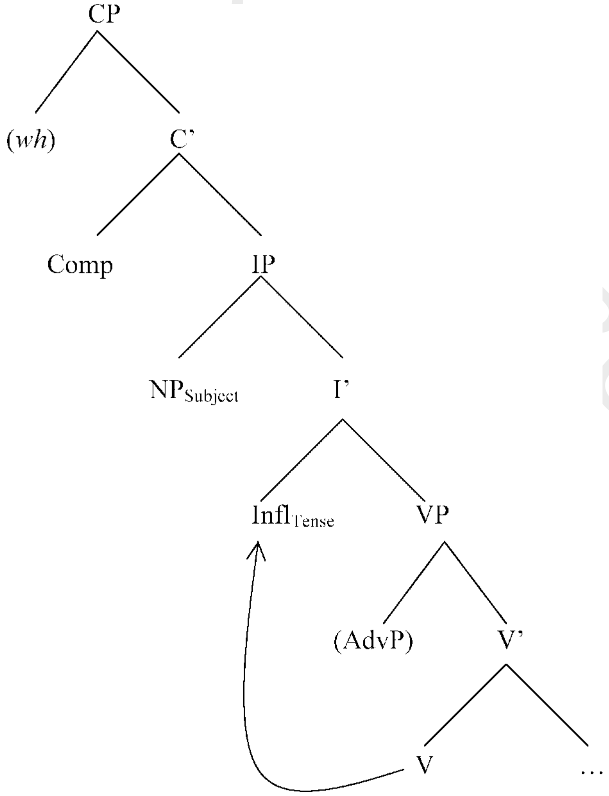


Figure 7.6 Verb-movement to Infl(ection) in Early Modern English.

it was not blocked (7.6c). These facts predict that the loss of Verb-to-Infl movement should occur generally across all sentence types.

(7.6) a. Where doth the grene knyght holde hym?
b. . . . bycause the nobylyte ther commynly dothe not exercyse them in the studys therof.
c. Me thinke I doe heare a good manerly Begger at the doore . . .
(Kroch 1989)

We can test this prediction by examining the distribution of the incoming form in different sentence types. Figure 7.7 shows the frequencies of *do* in five sentence types between 1400 and 1700 (Ellegård 1953): negative declaratives (with *not*) (7.5b); negative questions (with *not*) (7.7a); (affirmative) adverbial (with *where*, *when*, *why*) or *yes/no* questions that are transitive (7.6a); (affirmative) adverbial or *yes/no* questions that are intransitive (7.7b); and (affirmative) *wh*-object questions (7.7c).

(7.7) a. Did I not give it thee? (Warner 2005: 272)
b. Unhappy Gaveston, whither goest thou now?
(Marlowe, *Edward II* II.5)
c. What do you read, my lord? (Shakespeare, *Hamlet* II.2)

The curves in Figure 7.7 all appear to increase at different rates according to sentence type, apparently contradicting the constant rate hypothesis.

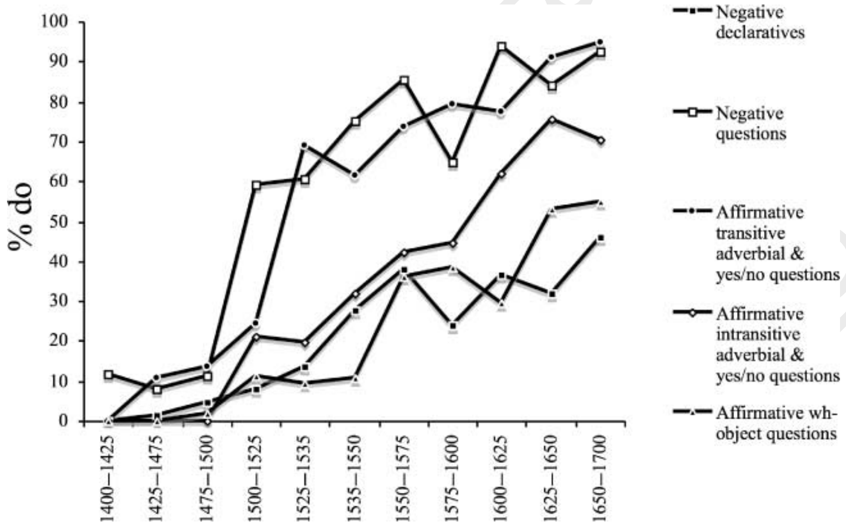


Figure 7.7 Frequency of periphrastic *do* in five linguistic contexts, 1400–1700 (adapted from Kroch 1989 and Ellegård 1953).

However, applying the logistic transformation to the curves in Figure 7.7, as shown in Table 7.2, we see that the slopes of all five curves are very close to an average value of 3.7. A statistical test shows that none of the variation around the average is significant, suggesting that the underlying rates of change are the same in all sentence types, with any deviations due to random fluctuation. Kroch uses additional evidence from the placement of adverbials such as *never* with respect to tensed verbs (7.8a vs. 7.8b) to support the idea that the spread of *do* is correlated with the general loss of Verb-to-Infl movement.

(7.8) a. Quene Ester looked never with swich an eye.
(Chaucer, *The Merchant's Tale*, line 1744).
b. Quene Ester never looked with swich an eye.
(Kroch 1989: 226)

However, one sentence type is not included in Figure 7.7: affirmative declaratives (7.6c). As Figure 7.8 shows, the curve of the frequency for this context looks very different from those in Figure 7.7. Kroch's application of the logistic transformation to the curve for affirmative declaratives reveals a slope of 2.82, significantly different from that of the other curves. Thus, the distribution of periphrastic *do* in affirmative declaratives constitutes a counterexample to the constant rate hypothesis. Kroch argues that the difference between affirmative declaratives and other contexts can be explained by positing not only variation between *do* and Verb-to-Infl movement, but also with a third variant, affix hopping, in which the tense-marking is lowered from Infl onto the Verb. For example, the sentence in (7.8a) is underlyingly something like (7.9a), but could surface in one of three forms (7.9b–d).

(7.9) a. Quene Ester Infl$_{[+\text{Past}]}$ [never look with swich an eye]$_{\text{VP}}$
b. *Verb-to-Infl Movement*
Quene Ester looked$_{[+\text{Past}]}$ [never ______ with swich an eye]$_{\text{VP}}$

Table 7.2 Comparison of slopes and intercepts for five linguistic contexts in the historical development of do-support in English, 1400–1700 (adapted from Kroch 1989).

	Slope (*s*)	**Intercept** (*k*)
Negative declaratives	3.74	−8.33
Negative questions	3.45	−5.57
Affirmative transitive adverbial and *yes/no* questions	3.62	−6.58
Affirmative intransitive adverbial and *yes/no* questions	3.77	−8.08
Affirmative *wh*-object questions	4.01	−9.26
Average:	3.7	

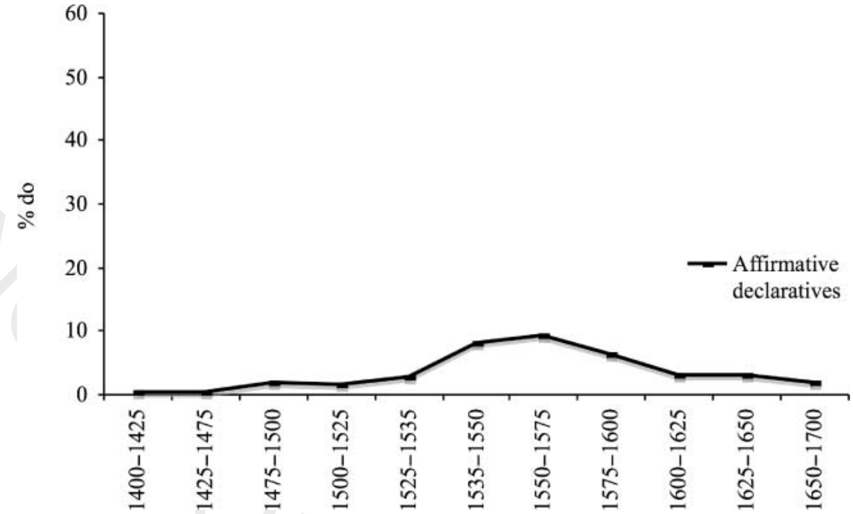


Figure 7.8 Frequency of periphrastic *do* in affirmative declarative contexts, 1400–1700 (adapted from Kroch 1989 and Ellegård 1953).

c. *Periphrastic* do
Quene Ester did$_{\text{[+Past]}}$ [never look with swich an eye]$_{\text{VP}}$
d. *Affix hopping*
Quene Ester ______ [never looked$_{\text{[+Past]}}$ with swich an eye]$_{\text{VP}}$

Under this view, the overall rate of periphrastic *do* with affirmative declaratives is lowered due to competition with another alternative to Verb-to-Infl movement, affix hopping. After 1575, affix hopping became the preferred form in this context, as seen in Modern English, where only (7.9d) is possible. Thus, what appears to be a unitary variable process actually involves interrelated variables in different linguistic contexts, featuring competition between different variants.

7.3.2 Grammaticalization

An alternative model of language change is afforded by the study of grammaticalization (Bybee et al. 1994), a type of language change in which forms from one part of the linguistic system take on functions in other parts of the system. The origin of this term is normally attributed to Meillet (1912: 131), who referred to it as “le passage d’un mot autonome au role d’élément grammatical . . . l’attribution du caractère grammatical à un mot jadis autonome” [“the passage of an autonomous word to the role of a grammatical element . . . the attribution of grammatical characteristics to a formerly autonomous word”, my translation]. In the classic case of grammaticalization, a lexical form takes on grammatical functions, a process most clearly exemplified in the cross-linguistically common development of future markers from the verb *go*. As we have seen, this happened in English, French, Spanish and Portuguese, as well as in other languages.

Subsequent studies have widened the definition of grammaticalization to include not only the movement of lexical forms into grammatical functions but also the movement of already-grammatical forms into other areas of grammar, such as syntactic forms into morphology or discourse markers. For example, the synthetic future in French developed from a periphrastic construction, in which an auxiliary verb (from Latin *habere* "to have (to)") gradually fused to the main verb (7.10). In English, erstwhile main clauses occurring without complementizer *that*, such as *I think* and *you know*, have developed into discourse markers that can occur not only in main-clause position but also at various positions in the sentence (7.11).

(7.10) *Grammaticalization: syntax* → morphology
Latin *cantare habeo* "to-sing I-have" > French *je chanterai* "I will sing"

(7.11) *Grammaticalization: syntax* → discourse marker
a. You know, I do believe in things happen for a reason. (TO5: 247)
b. They would have to, you know, be happy for me. (TO5: 130)
c. As long as they'd be careful with it, you know. (TO5: 175)

Studies of grammaticalization (Bybee et al. 1994, Hopper 1991) have uncovered a number of operative principles, several of which make predictions that can be translated into a variationist approach to language change. Variation between linguistic forms is recognized by the principle of layering, which states that multiple forms may undergo grammaticalization in the same functional domain. Under this principle, we expect forms to co-vary within functional domains, though their patterns of distribution within that domain may be shaped by other principles. For example, the domain of future reference in English is occupied not only by a *go*-future but also by a modal future *will*, which grammaticalized from a verb of volition. Because of this principle, grammaticalization studies have an advantage over strictly modular views of the study of language change, in that they have no problem in accommodating variation.

The gradualness of language change in grammaticalization is manifested in several potentially contradictory principles. The principle of persistence (or retention) states that forms undergoing grammaticalization retain semantic nuances inherited from their lexical sources. On the basis of this principle, we expect such forms to exhibit patterns of distribution originally associated with their lexical source, at least in the initial stages of grammaticalization. For example, we expect early stages of grammaticalization of the *go*-future to exhibit a greater tendency to association with verbs of motion and volitional subjects, since these properties are in keeping with the lexical meaning of *go*. As we saw with the results for Brazilian Portuguese *a gente* and English *be like* in Chapter 6, there is evidence for the persistence of lexical meaning in current patterns of grammatical variation.

However, other interrelated principles predict the progressive erosion of structure and meaning of forms undergoing grammaticalization. The principle of semantic bleaching (or desemanticization) states that forms gradually lose lexical-semantic content over time (although they may gain other types of semantic content), the principle of syntactic generalization states that forms become less restricted in the syntactic contexts in which they can occur, and the principle of erosion (or phonetic reduction) states that forms lose phonetic structure through processes such as contraction, coalescence and deletion. At the same time, forms also lose categorial information (for example, changing from nouns and verbs to prepositions, auxiliaries and other closed-class constituents). Taken together, these principles predict that the linguistic conditioning of the initial stages, associated with the lexical source, gradually weakens over time, and the form undergoing grammaticalization becomes phonetically and semantically reduced and less restricted semantically and syntactically. Certainly for grammaticalized variants of the future in English, these principles are operative. Both *going to* and *will* display phonetic reduction in their highly frequent contracted forms (*gonna* and *'ll*). Moreover, variationist studies of modern varieties of English have consistently failed to find an association between *going to* and verbs of motion or factors indicative of volition, arguing that these early-grammaticalization effects predicted by the principle of retention have largely dissipated.

Although many variationist studies assume grammaticalization in order to explain the conditioning of variation, there are actually very few variationist studies of grammaticalization that incorporate a comparative dimension across different time periods, either diachronically (real time) or synchronically (apparent time). As an example of a diachronic study, we consider Poplack and Malvar's (2007) analysis of the development of the *go*-future in Brazilian Portuguese (BP). This variant of future temporal reference, which was virtually nonexistent in BP before the nineteenth century, became the preferred option in the twentieth century, especially in colloquial speech. Using data from popular plays, Poplack and Malvar examine the distribution of different variants of future temporal reference in the nineteenth and twentieth centuries, comparing

Table 7.3 Factors contributing to the occurrence of the *go*-future in Brazilian Portuguese, in two time periods (Poplack & Malvar 2007).

	19th Century		20th Century			
			Plays		Speech	
Total N:	492		474		662	
Input:	.15		.81		.93	
Sentence Type						
Declarative	.59		[ ]		[ ]	
Negative	.10		[ ]		[ ]	
Interrogative	.31		[ ]		[ ]	
Range:		*49*				
Contingency						
Contingent	[ ]		.27		.13	
Assumed	[ ]		.52		.55	
Range:				*25*		*42*
Verb Type						
Non-motion	[ ]		.52		[ ]	
Motion	[ ]		.29		[ ]	
Range:				*23*		
Temporal Distance						
Distal	.36		[ ]		[ ]	
Proximal	.79		[ ]		[ ]	
Range:		*43*				
Grammatical Person/Animacy						
1st Animate	[ ]		.39		[ ]	
2nd Animate	[ ]		.72		[ ]	
3rd Animate	[ ]		.54		[ ]	
3rd Inanimate	[ ]		.50		[ ]	
Range:				*33*		
Adverbial Specification						
Non-specific	.15		.33		.43	
No adverbial	.62		.55		.58	
Specific	.27		.33		.20	
Range:		*47*		*22*		*38*

Factors not selected as significant: Type of clause, Presence of clitics.

these data with recorded speech from the twentieth century. Table 7.3 shows their results for separate variable rule analyses of the *go*-future by time period and genre. In the nineteenth century, the strongest effect is sentence type, with declaratives favoring the *go*-future and other sentence types, especially negatives, disfavoring. Temporal distance is also important, with proximal contexts favoring the *go*-future. However, both of these effects disappear in the twentieth century, where contingency is more important, an effect not significant in the nineteenth century, and is stronger in spoken than in written genres in the twentieth century. The only consistent effect is adverbial specification, with absence of an adverbial favoring the *go*-future. If we view written genres as more conservative than speech, we can view the results in Table 7.3 as revealing a trajectory of grammaticalization for the *go*-future in Brazilian Portuguese. In the initial stages of its grammaticalization, in keeping with the principle of retention, the *go*-future was associated with proximity, a reading that is consonant with the meaning of "going" more generally. By the twentieth century, however, this reading had disappeared, in keeping with the principle of semantic bleaching. Thus, the patterns of distribution shown in Table 7.3 provide a real-time illustration of grammaticalization over a 200-year period.

As another example of a diachronic variationist study of grammaticalization, we consider Torres Cacoullos's (in press) study of the development of the progressive in Spanish. As the examples in (7.12) show, variation between the progressive and simple present has existed since the fifteenth century.

(7.12) a. Está devaneando entre sueños. (15th C, Celestina, VIII)
is-3SG raving between dreams
"He is raving in his sleep."
b. Hijo, déxala dezir, que devanea; (15th C, Celestina, IX)
son, let-her talk, that (she) rave-3SG
"Son, let her talk, she is raving."
(Torres Cacoullos, in press)

The results of her analysis for three time periods (twelfth to fifteenth century, seventeenth century and nineteenth century) are shown in Table 7.4. In the twelfth to fifteenth century, the favoring effect of a co-occurring locative stems from the origins of the progressive in locative constructions. Although there is an aspectual effect, with limited duration favoring the progressive, its range is the same as that of other factor groups. In the seventeenth century, aspect becomes the most important factor group, stativity emerges as a significant effect, and the effect of locatives weakens. In the nineteenth century, the effect of locatives loses statistical significance. Thus, the principle of bleaching is revealed in the progressive weakening of locatives, while there is a gradual emergence of aspectual effects, with the progressive preferred in dynamic events of limited duration. This analysis provides another illustration of the trajectory of change of a grammaticalizing feature over the centuries.

An example of a synchronic (apparent-time) study of grammaticalization from a variationist perspective is provided by Tagliamonte and D'Arcy's (2007) study of the development of quotative *be like* in Toronto English, which we discussed in Chapter 6. This option for quoting reported speech has shown a rapid development, first reported

Table 7.4 Factors contributing to the occurrence of the progressive (vs. simple present) in three stages of Spanish (adapted from Torres Cacoullos, in press).

	12th–15th C		17th C		19th C	
Total N:	493		676		853	
Locative co-occurrence						
Present	.72		.71		[ ]	
Absent	.48		.48		[ ]	
Range:		*24*		*23*		
Aspect						
Limited duration	.62		.79		.69	
Extended duration	.38		.14		.17	
Range:		*24*		*65*		*52*
Polarity—Mode						
Affirmative declarative	.54		.57		.56	
Negative interrogative	.31		.15		.27	
Range:		*23*		*42*		*29*
Subject type + position						
Postverbal full NP	.70		[ ]		[ ]	
All others	.48		[ ]		[ ]	
Range:		*22*				
Transitivity						
Transitive	.61		[ ]		[ ]	
Intransitive	.45		[ ]		[ ]	
Range:		*16*				
Stativity						
Dynamic predicate	[ ]		.60		.56	
Stative predicate	[ ]		.27		.33	
Range:				*33*		*23*
Temporal co-occurrence						
Present	[ ]		[ ]		.64	
Absent	[ ]		[ ]		.18	
Range:						*16*

in the early 1980s and now the preferred option among young speakers. Tagliamonte and D'Arcy make use of the distribution in apparent time to adduce a grammaticalization path for *be like*, dividing their speakers into three age-groups: 30–39 (born 1965–1974), 20–29 (born 1975–1984) and 17–19 (born 1985–1987). The 30–39-year-olds can be considered to be among the earliest users of *be like*, and therefore represent an early stage of grammaticalization. Table 7.5, which shows the independent variable rule analyses of *be like* for each age group, reveals a three-stage model of the grammaticalization of this form. At Stage 1, *be like* is favored by inner thought and present tense, whether present temporal reference or historical present. In Stages 2 and 3, representing generations of speakers who have carried the

Table 7.5 Contribution of factors to the use of quotative *be like* in Toronto English in three age groups (adapted from Tagliamonte & D'Arcy 2007).

		Stage 1	**Stage 2**	**Stage 3**
		30–39 years	20–29 years	17–19 years
	Input:	.31	.72	.82
	Total N:	524	1,138	1,992
Tense				
Historical Present		.74	.73	.67
Present		.68	.50	.44
Past		.39	.34	.32
	Range:	*35*	*39*	*35*
Content				
Inner Thought		.70	.55	.54
Direct speech		.41	.48	.49
	Range:	*29*	*7*	*5*
Person				
First		.51	.56	.55
Third		.49	.44	.45
	Range:	*2*	*12*	*10*
Sex				
Male		.48	.52	.56
Female		.53	.47	.33
	Range:	*5*	*5*	*23*

change forward, the effect of content has weakened to the least important constraint. The tense effect is most interesting, since in Stages 2 and 3 the favoring effect of present tense more generally changes to that of historical present more specifically. This narrowing of the favoring effect suggests that *be like* has taken on more of a narrative function at later stages. If we can assume that speakers do not change the pragmatic structure of reported speech across the course of their lifespan, we can take the synchronic conditioning of variation by age group in Table 7.5 to reveal the path of grammaticalization for this discourse feature.

7.4 Conclusion

In this chapter, we applied the principles of variationist analysis to the study of language change. The variationist approach can accommodate the apparent gradualness of language change using linguistic conditioning to infer changes in the linguistic system, while other approaches face the problem of accounting for gradualness and incrementation. We invoked Weinreich, Labov and Herzog's (1968) problems for the study of language change—Constraints, Transition, Embedding (linguistic and social), Evaluation, Actuation—with a focus on Transition and (linguistic) Embedding. We contrasted studies in real time and apparent time and the advantages and disadvantages of each. Real-time studies are the only definitive way of studying language change, but they face problems of logistics. Apparent-time studies allow for more feasible synchronic study of language change but their effects may be confounded by age-grading and communal change.

We used quantitative distributions to test models of language change that have been proposed to address the questions of whether language change is gradual or abrupt and whether changes spread without exception or are conditioned by the linguistic context. Bailey's wave model relies on the S-shaped curve of language change and differential conditioning across linguistic contexts. Kroch argues against the wave model, proposing instead that changes begin differently in each context and proceed at a constant rate across all contexts. Kroch supports the constant rate hypothesis through a number of historical examples. Apparent counterexamples to the hypothesis may be explained by appealing to re-analysis and/or multiple variable contexts. The study of grammaticalization affords an alternative model of language change. The principles of layering, persistence and semantic bleaching make predictions about the expected conditioning effects of the lexical source of the construction undergoing grammaticalization and the pragmatic meaning it takes on as part of its new grammatical function. Variationist studies in real time and apparent time show how the principles of grammaticalization can be operationalized to reveal trajectories of change. Reconciling these different views of language change remains an ongoing challenge for variationist research. Such differences may have more to do with methodological or analytic differences rather than differences of interpretation. For example, studies that support the constant rate hypothesis tend to examine much smaller numbers of factors than do studies of grammaticalization within the variationist paradigm, and studies of grammaticalization tend to focus on the development of forms with discourse-pragmatic functions rather than the more abstract structural changes examined by Kroch.

The recognition of inherent variability and the use of conditioning by language-internal factors to infer linguistic systems has been used in this chapter to examine the nature of language change, through a comparison of rates and conditioning across time periods. In the next chapter, we make use of similar types of comparative analysis to resolve issues when more than one linguistic system is involved.

7.5 Further Reading

Bailey, Charles-James N. 1973. *Variation and Linguistic Theory*. Arlington, VA: Center for Applied Linguistics.

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1: 199–244.

Sankoff, Gillian. 2006. Age: Apparent time and real time. In Keith Brown (ed.) *The Encyclopedia of Languages and Linguistics, Vol. 1*. Cambridge: Elsevier, 110–16.

Weinreich, Uriel, William Labov and Marvin I. Herzog. 1968. Empirical foundations for a theory of language change. In W. Lehmann & I. Malkiel (eds.), *Directions for Historical Linguistics*. Austin: University of Texas Press, 95–195.

8 Language Contact

8.0 Introduction

Previous chapters established the principles of variationist analysis, which is concerned not only with whether a linguistic feature is present or absent, but also the relative frequency of its occurrence and, more importantly, the linguistic factors conditioning its occurrence. Indeed, variationist linguistics takes the linguistic conditioning of variation to reflect the linguistic system. Under this approach, we use the hierarchy of conditioning by language-internal factors to infer the linguistic system. By comparing the linguistic conditioning of variation between different varieties of language, we can determine the extent to which those varieties share a linguistic system.

The last chapter demonstrated the application of variationist analysis to the study of language change. Using the linguistic conditioning to infer the linguistic system, we assume that any changes in linguistic conditioning can be taken to indicate the presence of change. Comparing the linguistic conditioning of incoming forms across different time periods allows us to test competing models of language change.

In this chapter, we demonstrate the application of variationist analysis to several issues in the study of language contact. When two or more languages (or varieties of the same language) come into contact (more precisely, when speakers of different languages come into contact), various linguistic outcomes are possible (see e.g. Gardner-Chloros 2009 for an overview). At the very least, speakers may alternate between languages through "code-switching", either intersententially (across sentences), as in (8.1a), or intrasententially (within the sentence), as in (8.1b). Speakers may also borrow lexical items from one language into another, as in (8.2).

(8.1) a. *Code-Switching (Intersentential): English* → French
She kept me there for about ten minutes, 'til the man behind me says, "*C'est donc ridicule!*" ["This is ridiculous!"] (QC6: 1600)

b. *Code-Switching (Intrasentential): English* → Spanish
And from there I went to live *pa' muchos sitios* [in many places].
(Poplack 1980b)

(8.2) *Lexical borrowing: French* → English
So the way it worked with her Master's, it was like two months in school, ten months *stage* [[staʒ] "work term"], which was like a paid *stage*. And two months in school again. (MQ212: 370)

Code-switching and lexical borrowing are common manifestations of language contact, but they do not involve a great deal of interaction between the linguistic systems. Intrasentential code-switching requires a high degree of fluency in both languages (Poplack 1980b) precisely because code-switches must satisfy the structural requirements of both languages simultaneously. Similarly, words borrowed from one language into another tend to be integrated into the recipient language, usually grammatically and, to different degrees, phonologically (Poplack, Sankoff & Miller 1988).

However, there are other linguistic manifestations of language contact that do entail interaction between the structures of the two languages. Elements or features of one language may be transferred from one language to the other, under a process variously referred to as "interference", "transfer" or "structural borrowing". The languages in contact may gradually come to share a single linguistic system ("convergence"). At the most extreme end of the spectrum of linguistic consequences, an entirely new linguistic system may emerge, resulting in a "mixed language", one that combines elements of both languages, or a "pidgin" or "creole", whose features may stem from multiple sources. In fact, it is normally expected that languages in contact for long periods of time will necessarily influence each other (see Thomason and Kaufman 1988 for an overview).

The studies reviewed in this chapter focus on the linguistic consequences of language contact in which the linguistic systems are hypothesized to interact. Although lexical borrowing and code-switching are not without interest, the study of these processes raises issues that require different methodological and analytical considerations than those we have discussed (see Poplack 1993, Poplack & Meechan 1998). Given the lack of commonly agreed-upon principles for distinguishing borrowing and code-switching (see Gardner-Chloros 2009 and Muysken 2002), a discussion of the issues involved would be beyond the scope of this book. The focus of this chapter is the use of variationist analysis to resolve questions of system membership in situations of language contact; specifically, to test competing hypotheses about the linguistic consequences.

8.1 Second Language Acquisition

We begin by considering the linguistic consequences of acquiring a second language in adulthood, which is a rather different matter from acquiring a first (or second) language in childhood. Most linguists, whether or not they accept the existence of some sort of innate language acquisition device in childhood, would agree that language acquisition shuts down, or at least becomes less active, after puberty. As a result, adults are thought not to have access to the same language-learning cognitive resources that are available to children. Instead, adults may rely on features of their first language and/or general learning strategies not specific to language in building a "bridge" to the second language. This bridge is normally referred to by the term "interlanguage" (Selinker 1972).

Studies of adult second language acquisition, which take the learner's interlanguage as their object of study, note the high degree of variability that characterizes this process. This variability has a number of possible sources: language-learning errors, features transferred from the learner's first language, or universal linguistic strategies. Additionally, though not always recognized, native speakers of the target language also exhibit variable linguistic behavior, constituting an additional source of variation in the learner's language. The variationist approach is ideally suited to locating the sources of variability in second language acquisition, making use of linguistic conditioning as a tool for resolving system membership. By comparing the linguistic conditioning of variation observed in second-language speech with that of native-speaker speech, we can determine whether the two are parallel, or whether the variability can be attributed to other sources.

As an example, consider Bayley's (1996) study of variable (t/d)-deletion in the English of Chinese first-language speakers. As we have seen, this well-studied variable has been shown to be conditioned by several phonological and grammatical factors in native-speaker English. Comparing (t/d)-deletion in second-language English and native-speaker English will allow us to determine whether the source of variation stems from variability in the target language, considerations relevant to the speakers' first language, or second language learning strategies.

Table 8.1 shows Bayley's variable rule analysis of factor groups contributing to deletion in these second language English speakers. The same factor groups are selected as significant as for native speakers: the preceding and following phonological environment and the morphological status of the [t]/[d]. Moreover, the phonological conditioning of (t/d)-deletion in second language English is very similar to that of native speakers. Deletion is more likely the more features are shared between the preceding segment and the [t]/[d], and if the following segment is a consonant.

Table 8.1 Factors contributing to (t/d)-deletion in the English of 20 Chinese L1 speakers (adapted from combined data in Bayley 1996: 104).

	Total N:	3,170	
	Input:	.13	
Voicing (α_##)			
[αvce]		.67	
[-αvce]		.33	
	Range:		*34*
Preceding Segment			
Obstruent/Nasal		.64	
Liquid		.36	
	Range:		*28*
Morphological Status			
Past tense (preterit)		.66	
Past participle		.47	
Monomorpheme		.46	
Semiweak verb		.39	
	Range:		*27*
Following segment			
Obstruent/Liquid		.60	
Glide		.55	
Pause		.45	
Vowel		.40	
	Range:		*20*

Not selected as significant: Syllable stress, Cluster length
Significant but not shown: Style, Social network, English proficiency

However, the effect of morphological status reveals an important difference between second language speakers and native speakers. In native speaker English, deletion tends to be favored with monomorphemic forms (*mist*, *pact*) and disfavored with past-tense forms (*missed*, *packed*), with semiweak verbs (*kept*, *left*) having intermediary effects. In contrast with these results, Table 8.1 shows a result for second language English speakers that goes in exactly the opposite direction: past-tense forms favor deletion (.66), while monomorphemic forms disfavor (.46). That this is clearly a semantic effect (tense, or temporal reference) rather than a structural effect (morphology) can be seen in the fact that past participles, which are morphologically identical to preterits, disfavor deletion (.47) as much as monomorphemic forms.

What explanations are possible for the effects observed in Table 8.1? Bayley points out that these second language speakers exhibit another type of variability, that of marking past tense morphologically, on both regular and irregular verbs: that is, in addition to variably deleting word-final [t]/[d], they also variably show stem changes on irregular verbs (for example, *sing/sang*). Thus, the results shown in Table 7.1 may result from the confluence of two variables, one grammatical (past marking) and one phonological (word-final (t/d)-deletion). Unmarked regular past tense verbs may result from either of these variable processes.

Additional explanations for the pattern in Table 8.1 may appeal to the first language of the learners, either generally (i.e. all second language learners of English) or specifically (i.e. Chinese learners of English). The first possibility is ruled out by a comparison of the ranking of morphological factors in different varieties of English, shown in Figure 8.1. In contrast with Chicano and Tejano varieties of English (which may be Spanish-influenced), only Chinese second language English favors deletion with past tense. Unlike English and Spanish, which both have a morphologically marked category "tense" (past/non-past), Chinese morphologically marks verbs according to aspect (completed/non-completed). Thus, a better explanation for the results for morphological status shown in Table 8.1 is variable past-marking involving influence from the first language.

8.2 Convergence

When languages are in contact for long periods of time, it is normally expected that they will gradually converge: that is, they will come to resemble each other more and more. The classic case of convergence is the Balkan *Sprachbund* ("linguistic federation"), in which the historically structurally different languages of the Balkan peninsula (Greek, Albanian, Bulgarian and Romanian) have gradually converged in their linguistic structures after centuries of co-existence. The languages are said to be so similar that sentences are virtually identical, differentiated only by lexical choices, as shown in the example sentence in (8.3).

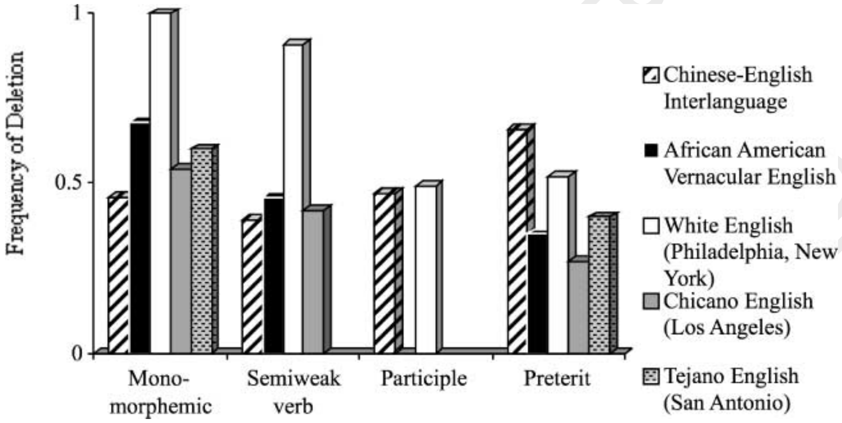


Figure 8.1 Contribution of morphological status to (t/d)-deletion in Chinese-English interlanguage and native English varieties (adapted from Bayley 1996: 109).

(8.3)	Albanian	*due*	*te*	*shkue*
	Bulgarian	*iskam*	*da*	*otida*
	Romanian	*veau*	*sa*	*plec*
	Greek	*thelo*	*na*	*pao*
		I-want	that	I-leave
		"I want to go"		

(Appel & Muysken 1987: 155)

Another example of convergence is Gumperz and Wilson's (1971) study of the village of Kupwar, India, in which the historically structurally different languages Urdu, Marathi and Kannada have over the centuries achieved a *Sprachbund*-like level of intertranslatability, as shown in (8.4).

(8.4)	Kupwar Urdu	*pala*	*jəra*	*kat*	*ke*	*le*	*ke*	*a*	*Ø*	*ya*
	Kupwar Marathi	*pala*	*jəra*	*kap*	*un*	*ghe*	*un*	*a*	*l*	*o*
	Kupwar Kannada	*tapla*	*jəra*	*khod*	*i*	*təgond*	*i*	*bə*	*Ø*	*yn*
		greens	a-little	having	cut	having	taken		I	came
		"I cut some greens and brought them."								

(Gumperz & Wilson 1971)

Situations such as the Balkans and Kupwar are frequently offered as evidence that languages in contact over long periods of time will gradually resemble each other in their linguistic systems, but the evidence provided in each case relies on categorical data taken out of context. As we saw in the previous chapter, all change implies variation, which suggests that there must have been a period of variation that led to these situations. Although convergence would seem an ideal area of research for the variationist approach, there are few such studies.

Surprisingly, the few variationist studies of situations of contact where convergence is expected provide evidence that speakers <u>can</u> maintain distinct linguistic systems even in situations of intensive, long-term contact. An early example is Rickford's (1985) study of two speakers of varieties of English in the Sea Islands in South Carolina, where African Americans and white Americans have co-existed for centuries. Rickford's examination of the speech of two Sea Islanders, Mrs. Queen (an African American woman) and Mr. King (a white American man), shows that they share many phonological features characteristic of both Southern American English and African American English. However, although a number of grammatical variants are present in the English of both speakers, the quantitative distributions of these features is quite different for the two speakers. As Table 8.2 shows, the variants of plural marking and passive formation are shared by both speakers (except for plural *-dem* for Mr. King), but each speaker shows markedly different preferences. For plural marking, Mr. King clearly prefers the suffix *-s* (94%), while Mrs. Queen tends not to mark plural nouns (76%). Similarly, while Mr. King favors

Table 8.2 Overall distribution of plural marking and passive formation in the speech of two Sea Islanders (South Carolina) (adapted from Rickford 1985).

	Mrs. Queen (African American)	**Mr. King** (White)
Plural Marking		
-s	23%	**94%**
-*Ø*	**76%**	6%
-dem	1%	0%
Passive Formation		
be	6%	**64%**
get	23%	23%
Unmarked	**71%**	13%

be passives (64%), Mrs. Queen prefers unmarked passives (71%). These results demonstrate that speakers of linguistic varieties in contact for long periods of time that share features can nevertheless show distributional preferences for different variants.

In addition to comparing the overall distribution of forms across speakers, as in Table 8.2, we can compare the linguistic conditioning across varieties as a means of determining the results of language contact. We return to Meyerhoff and Walker's (2007) study of the island of Bequia (St. Vincent and the Grenadines), which we discussed in Chapter 6. Bequia is characterized by a high degree of linguistic variation, in which communities on the island are in close contact but distinguished by different histories of settlement and socioeconomic environments. Meyerhoff and Walker concentrate on three communities: Hamilton, a predominantly African-descent community originating from a large plantation; Mount Pleasant, the traditional home of the British-descent population; and Paget Farm, an ethnically mixed fishing and whaling community on the south side of the island. The linguistic focus of this work is zero copula, which, as the examples in (8.5) show, is shared by speakers in all three communities. Rickford's study provided evidence that speakers in close contact can share features but differ in their overall distribution. Are the communities on Bequia similarly distinguished by different frequencies of zero copula and, more importantly, by different linguistic conditioning of zero copula?

(8.5) a. He Ø making speed, running. (BQ.H3: 217)
b. She Ø gon run to see. (BQ.M104: 3000)
c. He Ø there in Antigua. (BQ.P19: 731)

Table 8.3 shows Meyerhoff and Walker's variable rule analyses of the contribution of factors to zero copula in each of the three communities. As shown by the input values at the top of each column, each community

Table 8.3 Factors contributing to zero copula in three communities on Bequia (St Vincent and the Grenadines) (adapted from Meyerhoff & Walker 2007).

	Hamilton (Insertion Analysis)	Mt. Pleasant (Deletion Analysis)	Paget Farm (Insertion Analysis)
Total N:	1002	640	690
Input (p^0):	.386	.459	.250
Following grammatical category			
gonna	.90	.83	.96
Verb-*ing*	.82	.79	.84
Adjective	.64	.47	.54
PP	.38	.53	.42
NP	.16	.12	.14
Locative adverb	.08	.53	.54
Range:	*82*	*71*	***82***
Subject type + preceding segment			
NP, Vowel	.58	.87	.50
Pronoun	.53	.47	.50
NP, Consonant	.43	.55	.49
Range:	*15*	*40*	***1***

is characterized by different overall probabilities for the occurrence of zero copula: .386 for Hamilton, .459 for Mount Pleasant, and .25 for Paget Farm. As noted at the top of each column, the communities are additionally distinguished by different methods of calculating rates of zero copula (see Chapter 6): in Hamilton and Paget Farm, an insertion analysis provides the best fit to the data, while in Mount Pleasant a deletion analysis provides the best fit. This difference already provides evidence that the linguistic systems underlying the variation are different in one community.

In all three communities, the same factor groups are selected as significant (subject type and following grammatical category), but the effects are different. In Paget Farm, subject type, although significant, has minimal effect. In Hamilton, subject type is also significant, but appears to reflect phonological considerations, since preceding consonants disfavor zero and preceding vowels favor, regardless of whether the subject is a pronoun or NP. In Mount Pleasant, subject type is significant, but its effect is more clearly syntactic, with NP subjects favoring zero and pronouns disfavoring. For the following grammatical category, in all three communities there is a split in zero copula between its functions as an auxiliary with verbal predicates (*gonna*, V-*ing*), where it is strongly favored, and its function as a copula with non-verbal predicates (NP), where it is disfavored.

The biggest difference between communities lies in the behavior of following predicate adjectives. In Hamilton and Paget Farm, adjectives behave more like verbal predicates, in that they favor zero (especially in Hamilton), while in Mount Pleasant adjectives behave like non-verbal predicates, disfavoring zero. (The effects of the following locative adverb are complicated by homophony between the locative copula *deh* and the adverb *there*, which makes some tokens ambiguous.) Thus, despite the small size of the island, frequent contact between the communities, over 100 years of co-existence, and the presence of the same feature in all three communities, this analysis of the linguistic conditioning of zero copula demonstrates that these communities all have different linguistic systems.[1] However, these systems are distinguished in rather subtle ways that can be discerned only through quantitative analysis. These results show how the variationist method can be used to disentangle membership in linguistic systems that have co-existed for a long time and even share many features.

8.3 Pidgins and Creoles

In situations of language contact, the most extreme linguistic outcome is the emergence of an entirely new linguistic system. Although mixed languages in general have received a great deal of attention, undoubtedly the most attention has been paid to pidgins and creoles. Various theories have been adduced for the origins of their linguistic systems: contributions from the indigenous languages of the slaves or indentured laborers who created these languages ("substratist"), inheritances from the (usually) European languages of the conquerors ("superstratist"), and the role of linguistic processes innate to all humans ("universalist") (see Holm 1988 for an overview).

Importantly, both pidginization and creolization are diachronic processes that, like other types of language change, are characterized by variability. Since many of the theories of pidgin and creole genesis are couched in research traditions that have difficulty in recognizing variability, they often appeal to theories about the mixing of discrete linguistic systems or macaronic interference from the speakers' first languages (Bickerton 1975, 1981). As we have seen with other situations of variability, the variationist approach is ideally suited to deciding among competing hypotheses of the source of variability in the emergent linguistic systems of pidgins and creoles. Here we consider a few examples of studies that make use of quantitative modeling to decide among hypotheses.

First we examine Meyerhoff's (2000) analysis of null subjects in Bislama, an English-based creole spoken in Vanuatu. Table 8.4 shows the paradigm for subject-verb agreement in Bislama, using the verb *karem* "carry, bring" as an example. However, as Meyerhoff notes, null subjects variably occur in all persons and numbers, as shown in (8.6).

Table 8.4 Paradigm for finite verbs in Bislama: *karem* “carry, bring” (Meyerhoff 2000).

	Singular	Dual	Trial	Plural
1 (incl.)	—	*yumitu* karem	*yumitri* karem	*yumi* karem
1 (excl.)	*mi* karem	*mitufala* **i** karem	*mitrifala* **i** karem	*mifala* **i** karem
2	*yu* karem	*yutufela* **i** karem	*yutrifala* **i** karem	*yufala* **i** karem
3	*hem* i karem	*tufala* **i** karem	*trifala* **i** karem	*olgeta* **oli** karem

(8.6) a. *Denis hem i kam, Ø i blokem hem.* (S-95-7, Sevi)
Dennis 3sg *i* come, Ø *i* block 3sg
“Dennis came [and he] stopped her.”
b. *O, Ø talem se tangkiu tumas long hadwok blong hem.* (S-94-1, Timoti)
oh Ø say that thank-you too-much for labor of 3sg
“Oh, [I] said thank you very much for all her hard work.”

The form *i* is historically derived from the English pronoun *he*, but behaves differently in Bislama. Since many of the Austronesian languages spoken by the indentured laborers of the South Pacific who created Bislama contain a grammatical element that serves to indicate the predicate, *i* seems like a relexification of a grammatical category transferred from the Austronesian languages.

As Meyerhoff notes, there are two points in the paradigm in which *i* does not surface. In third person plural, *oli* occurs instead, and in first and second singular and first person inclusive. Meyerhoff provides evidence against a phonological rule assimilating *i* to the preceding vowel, which leaves two possible analyses of the status of *yu* and *mi*: predicate markers (like *i* and *oli*) or subject pronouns. The possibility that *yu* and *mi* are predicate markers is supported by apparent reduplication of forms, as in (8.7a), parallel to examples found in the third person (8.7b).

(8.7) a. *Mi mi kae.*
1sg 1sg eat
“I’m eating/I eat”
b. *Hem i go.*
3sg i go
“He’s going/He goes”

Meyerhoff tests these competing analyses by examining the behavior of null subjects in different discourse contexts. Subjects may surface as null or overt depending on the presence of coreferential elements in the preceding clause. The subject of the current clause may be the same as the subject or an element other than the subject in the preceding clause, or its referent may not occur in the preceding clause. In the first two options, the preceding coreferential element may surface as overt or null. Both of these considerations are relevant in conditioning the form of the subject in the current clause. We can therefore use this conditioning to determine whether the behavior of these ambiguous forms parallels that of forms which are clearly predicate markers, or those which are clearly pronouns. If *yu* and *mi* are pronouns, their conditioning by interclausal relations should parallel that of third person pronouns. This hypothesis is formulated as in (8.8).

(8.8) *Pronoun hypothesis*:

$$\frac{\text{1st/2nd person}}{\text{3rd person}}:\frac{\text{mi mi V}}{\text{NP, hem i V}}\text{ vs. }\frac{\text{mi V}}{\text{hem i V}}\text{ vs. }\frac{\text{Ø V}}{\text{Ø i V}}$$

However, if *yu* and *mi* are predicate markers, we expect the conditioning of singleton occurrences of *yu/mi* to parallel that of third singular forms with null subjects, and we expect the conditioning of reduplicated forms (e.g. *mi mi kae*) to parallel that of topicalized NPs (NP$_i$, *hem*$_i$ *i go*).

(8.9) *Agreement marker hypothesis*:

$$\frac{\text{1st/2nd person}}{\text{3rd person}}:\frac{\text{mi mi V}}{\text{hem i V}}\text{ vs. }\frac{\text{mi V}}{\text{Ø i V}}$$

Table 8.5 shows a series of pairwise comparisons of the conditioning subject types in first/second person and third person contexts, with a Pearson correlation test. Positive values indicate a positive correlation (i.e. similar conditioning), while negative values indicate a negative correlation (i.e. different conditioning). The numerical value indicates the strength of the correlation. The correlations in Table 8.5d and 8.5e are negative: that is, the conditioning is not similar between first/second person and third person. The correlations in Table 8.5a and 8.5b are positive, but low. The correlation for Table 8.5c is positive and provides the highest correlation value, showing that null subjects in first/second person are conditioned similarly to null third singular subjects. Therefore, we can reject the hypothesis that *yu* and *mi* are predicate markers, and they must therefore be pronouns. As Meyerhoff notes, the variability in null subjects is most robust where the verbal morphology is the most informative: that is, null subjects emerge as the norm in third person, rather than in first or second person, because the existence of an overt predicate marker in third person singular and plural makes subject/verb agreement the most informative. Thus, the emergence of null subject in Bislama results from the functional constraints of pragmatics.

In the remainder of this chapter, we consider further the question of

Table 8.5 Two-way comparisons of subject patterns according to interclausal relations (adapted from Meyerhoff 2000).

	Current subject = preceding subject		**Current subject = preceding non-subject**		**Current subject not in preceding clause**
	Preceding overt	**Preceding null**	**Preceding overt**	**Preceding null**	
a.					
mi mi	.428	.49	.72	.2	.6
NP, *hem i*	.366	.29	.48	.39	.75
					Pearson correlation: .449
b.					
mi	.572	.52	.29	.8	.41
hem i	.634	.71	.52	.61	.25
					Pearson correlation: .448
c.					
Ø	.457	.86	.43	.24	.5
Ø *i*	.389	.77	.47	.41	.29
					Pearson correlation: .772
d.					
mi mi	.428	.49	.72	.2	.6
hem i	.634	.71	.52	.61	.25
					Pearson correlation: –.449
e.					
mi	.543	.14	.57	.76	.5
Ø *i*	.389	.77	.47	.41	.29
					Pearson correlation: –.773

how grammatical features are developed in pidgins and creoles. In the classic scenario developed by Derek Bickerton (1981), tense in creoles follows a relative rather than absolute system. Rather than marking past/non-past, creoles mark anterior/non-anterior, a distinction that is sensitive to the stativity of the verb. In this scenario, zero-marked statives (8.10a) are interpreted as present, zero-marked non-statives (8.10b) are interpreted as past, statives marked with an anterior morpheme (such as *bin*) (8.10c) are interpreted as past, and non-statives marked with anterior (8.10d) are interpreted as past-before-past.

(8.10) a. *Mi sii Jan*
"I see John."
b. *Mi iit di mango*
"I ate the mango."
c. *Mi bin sii Jan*
"I saw John."
d. *Mi bin iit di mango*
"I had eaten the mango."

As Sankoff (1990) notes, this system sets up an opposition in which the presence of a form indicates a semantic feature and its absence indicates the opposite value of that feature: that is, the occurrence of *bin* is taken to indicate anteriority, and its absence indicates non-anteriority. However, a number of studies (summarized in Holm et al. 2000) have pointed out exceptions to Bickerton's system. Examining natural discourse in creole communities reveals a great deal of variability, in which *bin* sometimes does and sometimes does not occur in anterior contexts.

Nevertheless, Bickerton's system may represent a probabilistic tendency rather than a categorical rule. In other words, although there may not be a categorical association between *bin* and anteriority, there may be a greater tendency for *bin* to be associated with anterior contexts. If this is true, Bickerton's system lends itself well to the possibility of variationist analysis. To this end, Sankoff (1990) examines the distribution of the anterior marker *bin/ben* in two English-based creoles: Tok Pisin, spoken in Papua New Guinea, and Sranan, spoken in Surinam. As Table 8.6 shows, out of a total of 403 clauses in Tok Pisin, only three are marked with *bin*. While these three tokens do occur in one of the expected anterior contexts (past-before-past with non-statives), the fact that the majority of tokens in this category (14/17) are bare suggests that the absence of *bin* cannot be construed as indicating non-anteriority. In the Sranan texts, unambiguously anterior contexts are more difficult to identify, but even here fewer than 10 percent of tokens of stative verbs in past contexts (49/536), where Bickerton's system predicts *ben* to occur, are marked with *ben*. Thus, Sankoff concludes that an examination of the distribution of putative markers of anteriority provides little quantitative support for Bickerton's system of creole tense-marking.

Bickerton's proposal is couched in a broadly generativist approach to language that does not recognize gradualness and variability. In most versions of his theory of creolization, he envisions a one-generation

Table 8.6 Distribution of *bin/ben* according to predicate type in Tok Pisin and Sranan (adapted from Sankoff 1990).

	Tok Pisin		**Sranan**	
	N *bin*	Total N	N *ben*	Total N
Nonstatives				
Past	0	325	15	451
Past-before-past	3	17	?	?
Statives				
Past copula	—		6	9
Past nonpunctual	0	15	5	13
Past modals	0	6	10	24
Past statives	0	40	13	39
Total:	3	403	49	536

model of change. However, if we adopt the view of grammaticalization discussed in the previous chapter, creole tense markers may not develop abruptly but rather are gradually grammaticalized from lexical material. Under this view, the patterns in Table 8.6 may represent creoles at different stages in the grammaticalization of tense markers, in which *bin/ben* has not yet become the default marker of anteriority but is beginning to be preferentially associated with anteriority.

This line of inquiry is pursued in Poplack and Tagliamonte's (1996) analysis of an array of tense/aspect markers in Nigerian Pidgin English: *kom, don* and *bin*, as exemplified in (8.11).

(8.11) a. *wi kom drink am wit evriting wey i giv os* (4: 256)
"We drank it with everything he gave us."
b. *i don dai* (1: 013)
"He has died."
c. *a bin orijinali kom from Inglan* (1: 7–8)
"I originally came from England."
(adapted from Poplack and Tagliamonte 1996: 73)

Table 8.7 shows their variable rule analyses for these three markers. As the input values at the top of each column indicate, these markers (especially *bin*!) occur at very low frequencies. Most verbs with past temporal reference are unmarked. Nevertheless, when these markers do occur, each is preferentially associated with particular tense/aspect configurations in a probabilistic way: *kom* with sequential contexts (.70), *don* with anterior (.76) and non-remote (.63) contexts, and *bin* with anterior (.90) and remote (.58) contexts.

Poplack and Tagliamonte suggest that, if the input value can be taken as an indicator of relative position along a cline of grammaticalization, we should expect to see concomitant indicators of grammaticalization. Table 8.8 shows Poplack and Tagliamonte's (1996) measurements of other measures of grammaticalization: the overall frequency of each form, its frequency in its associated semantic context, its degree of phonological reduction (as measured by frequency of assimilation of the final nasal to the following segment) and the extent of the rigidification of each form in its syntactic position (as measured by the frequency with which open-class forms can intervene between the form and the verb, and the frequency with which the form occurs in the position immediately preceding the verb). The first three measures show a high degree of correlation: the more frequent a form is, the more frequent it is in its associated semantic context, and the more phonologically reduced it is. Moreover, the very low rate of open-class intervention (.5–3%) and the very high rate of occurrence immediately preceding the verb (96–100%) suggests that these forms have already achieved a high degree of grammaticalization in their structural constraints.

Table 8.7 Contribution of factors to three variants of past temporal reference in Nigerian Pidgin English (adapted from Poplack & Tagliamonte 1996: 81).

		kɔm	*dɔn*	*bin*
	Input:	.19	.07	.004
Temporal relationship				
Anterior		.20	.76	.90
Sequential		.70	.28	.21
Non-anterior		.41	.65	.63
	Range:	*50*	*48*	*69*
Temporal distance				
[+Remote]		[ ]	.44	.58
[–Remote]		[ ]	.63	.32
	Range:		*19*	*26*
Lexical stativity				
[+Stative]		.55	.48	[ ]
[–Stative]		.48	.51	[ ]
	Range:	*7*	*3*	
Temporal adverb				
Adverb present		.33	.36	.65
No adverb		.52	.51	.49
	Range:	*21*	*15*	*16*
Mark on preceding verb				
Same		.72	.82	.95
Different		.45	.49	.48
No mark		.41	.46	.51
	Range:	*31*	*36*	*37*
Negation				
Negative		.14	0	[ ]
Affirmative		.53	.50	[ ]
	Range:	*39*		

The results of this study provide further evidence for the utility of the variationist approach in elucidating the trajectory of change as forms grammaticalize in the process of creolization.

8.4 Conclusion

Research on language contact frequently notes the high degree of linguistic variability that results. This variability poses a problem for theories that rely on categorical approaches to the study of language. In contrast, the variationist approach recognizes that all language, whether or not found in situations of language contact, is inherently variable. As such, it is ideally suited to study not only monolingual speech communities but also communities in which a number of languages or language varieties co-exist. Specifically, the recognition that the linguistic

Table 8.8 Indices of grammaticalization for three variants of past marking in Nigerian Pidgin English (adapted from Poplack & Tagliamonte 1996).

	OVERALL FREQUENCY	FREQUENCY IN ASSOCIATED SEMANTIC CONTEXT	PHONOLOGICAL REDUCTION	RIGIDIFICATION OF SYNTACTIC POSITION	
			Consonant assimilation	**Open class intervention**	**Position preceding verb**
bin	2%	5% (anterior)	7%	3%	96%
dɔn	10%	15% (non-remote)	13%	3%	100%
kɔm	23%	38% (sequential)	44%	.5%	99%

conditioning of variation can be taken as an indication of the linguistic system allows us to compare this conditioning across varieties to test hypotheses about the linguistic outcomes of language contact.

In this chapter, we demonstrated the application of the variationist method to situations in which multiple linguistic systems are expected to interact, ranging from code-switching and borrowing, through second language acquisition and convergence to pidgins and creoles. Although variability in second language acquisition can be attributed to a number of sources (errors, language transfer, strategies of second language learning), we have shown how the variationist approach can be used to isolate these explanations. Similarly, in situations where languages in contact over long periods of time are expected to converge in structure, variationist analysis provides evidence for the maintenance of distinct linguistic systems. In the most extreme linguistic outcome of language contact, pidgins and creoles, various theories of origin (substratist, superstratist, universalist) have been proposed. Using quantitative modeling to test models of tense/aspect systems, we have shown that such systems are not categorical but rather reflect processes of grammaticalization that can be inferred from the quantitative patterning of forms. In each case, we provided a demonstration of the utility of the variationist approach in evaluating hypotheses of the linguistic consequences of language contact.

Up until now we have demonstrated how the principles of variationist analysis may be used to resolve issues in situations of contact and change, where membership in linguistic systems must be determined in order to decide among competing hypotheses. We have not considered the relationship between the variationist method and theories of language more generally, where no change or contact is assumed. In the next and final chapter, we consider this question.

8.5 Further Reading

Poplack, Shana. 1993. Variation theory and language contact: concepts, methods, and data. In Dennis R. Preston (ed.), *American Dialect Research*. Amsterdam: Benjamins, 251–86.

Poplack, Shana and Marjory Meechan. 1998. Introduction: How languages fit together in codemixing. *International Journal of Bilingualism* 2: 127–38.

Preston, Dennis R. 1989. *Sociolinguistics and Second Language Acquisition*. Oxford: Blackwell.

9 Conclusion

9.0 Introduction

Previous chapters have outlined the study of linguistic variation and its application to the questions of language contact and change. We began by defining linguistic variation and explored variation at different levels of the linguistic system. We established the methodological and analytical principles of the variationist approach to the study of language, considering a number of theoretical and methodological issues in conducting variationist analysis at the level of sound systems and grammatical systems. We applied the variationist method to various issues in the study of language change and language contact.

Throughout these chapters, we have often invoked linguistic theory in explanations of variation, operationalizing predictions of different theories as factors within a variationist analysis. However, we have not considered the other side of the relationship between variation and theory. What can variationist linguistics contribute to linguistic theory? How can variation be accommodated in linguistic theory, especially since theoretical linguistics normally operates on the assumption that linguistic behavior is categorical. In this chapter, we consider ways in which the study of variation and linguistic theory can be reconciled. We begin by reviewing treatments of variation in linguistic theory, focusing on the use of Optimality Theory in phonological theory and the Minimalist Program in grammatical theory. We conclude with a variationist perspective on these treatments.

9.1 Variation and Linguistic Theory

In previous chapters, we have been concerned with establishing the principles of variationist analysis and applying them to situations in which linguistic systems are hypothesized to change or to come into contact with other linguistic systems. Such situations are external to the linguistic system, as are other contexts in which variationist analysis has been applied, such as socio-symbolic and stylistic differences (see Eckert 1999 and Eckert and Rickford 2001). Because the variationist method has primarily been used to resolve issues originating from outside of the linguistic system, there is a widely held belief in linguistics that the study of linguistic variation is appropriate for sociolinguistics but stands apart from other, categorical approaches to linguistic analysis.

Based on the results of some of the studies reported in previous chapters, in which functional considerations have been found to condition the variation, we might be tempted to do away with formal, autonomous theories of language altogether, and derive all linguistic behavior from functional constraints. In "usage-based" theories of language, such as Exemplar Theory (Bybee 2006; Pierrehumbert 2001), linguistic structure is entirely dependent on speaker usage. Under this view, language is an "emergent system" (Hopper 1998) that arises from the speaker's experience, based on considerations of interaction and frequency. In previous chapters, we have seen some evidence for the role of functional and interactional constraints on both phonic and grammatical variation. For example, lower rates of (t/d)-deletion when the [t] or [d] serves to mark past tense may result from the greater semantic load of the phonetic segment in this context (Chapter 5). Higher rates of null subject in third person in Bislama may occur because this is the most informative context (Chapter 8). However, we have also seen evidence in both phonic and grammatical variation for the effect of formal or structural constraints on variation. For example, contrary to functional predictions, Spanish (s)-deletion appears to operate across all the tokens within a noun phrase rather than preserving information about plurality (Chapter 5). The subject position in relative clauses is most disfavorable to zero relative in English (Chapter 6). While we cannot deny that frequency and function play a role in conditioning variation, we should not completely abandon the idea that language has formal structure.

As we saw in Chapter 2, although variation is often acknowledged in linguistic descriptions, in linguistic theory it tends to be viewed as a problem to be avoided or solved. This attitude stems in part from the theoretical division between competence and performance (Chomsky 1965). Performance is assumed to be marred by disfluency, hesitations, and speaker errors resulting from the physiology of language production and other language-external considerations, such as the speaker's mood and attention. Competence, representing the knowledge of language contained in the individual speaker's brain, is assumed to be more systematic and orderly, and not subject to extralinguistic influences. Under this view, the only proper object of linguistic study is competence. This view has led to a methodological bias in linguistic theory building against natural speech data and in favor of grammaticality judgments by native speakers, based on either elicitation or intuition. Leaving aside the question of whether native-speaker judgments might be a type of performance, also affected by extralinguistic considerations (see Schütze 1996), the results reported in previous chapters provide ample evidence against the assumption that performance is entirely unsystematic. In fact, variationist analysis, which allows us to test this assumption empirically, reveals a great deal of systematicity in the conditioning of linguistic variation by linguistic factors. Such systematicity is not expected if performance is entirely unrelated to competence, though it would be very surprising if performance did not (in part) reflect considerations of competence (a point made early in variationist linguistics by Cedergren and Sankoff 1974).

Once we acknowledge the fact of variation in language, how do we accommodate it in a theory of language? One solution is to push its operation outside of the linguistic system, viewing variation as the co-existence in a speech community of different categorical linguistic systems (lects) defined in language-external terms, such as socially (sociolects) or regionally (dialects). As we saw in Chapters 7 and 8, this view is often adopted (implicitly or explicitly) in studies of language change and pidgins and creoles. This view implies that if we could only control for all language-external factors (such as social group, region or style), we would end up with an invariant linguistic system. However, as we have seen, variation persists no matter how thinly we slice the data (socially, regionally, or even at the level of the individual speaker). In previous chapters, we saw a great deal of evidence for the conditioning of linguistic variation by language-internal factors. While some of these factors may be traced to universal articulatory or functional considerations, in many cases we have been able to adduce evidence for phonological, morphological, syntactic and lexical effects, all of which reflect aspects of the linguistic system. The conditioning of variation by linguistic constraints renders it highly unlikely that all variation is due to language-external considerations.

Even acknowledging the existence of variation at the level of the individual, we might still argue that it is an epiphenomenon of co-existing linguistic systems. Kroch (2001) argues that the examples of language change discussed in Chapter 7 constitute grammar competition, the gradual replacement of one (categorical) linguistic system by another over time. Grammar competition is analogous to code-switching between different languages in bilingual communities and speakers. While such an interpretation is possible, it raises problems on both theoretical and empirical grounds. First, if every variant of every variable represents an entirely different grammar, there would be an enormous amount of duplication among linguistic systems, presenting problems for the storage and processing of language in the speaker's brain. Second, it is difficult to know how to distinguish true variation from grammar competition on an empirical basis. Appealing to diagnostic features of bilingual code-switching is made difficult by the absence of agreed-upon constraints on code-switching (as mentioned at the beginning of Chapter 8). Therefore, while grammar competition remains a possible interpretation of variation, at present it lacks empirical verifiability.

Although one argument against grammar competition is the unnecessary duplication of information across co-existent linguistic systems, we could avoid this problem by assuming multiple "mini-grammars". Under this view, a linguistic system consists of an overarching superset of grammatical structure containing within it smaller alternating parts that differ from each other minimally. Speakers have only one grammar, but they may choose among competing mini-grammars within that grammar. This view is adopted by Henry (1995) in her analysis of a number of syntactic variables in Belfast English. For example, in addition to standard marking on third person singular verbs, Belfast English also allows verbal *–s* to appear on third person plural verbs with NP subjects, as in (9.1) (recall the examples (6.2–6.3)).

(9.1) a. These cars go/goes very fast.
b. The eggs are/is cracked.
(Henry 1995: 16)

As Henry notes, since most of the grammatical structure of Standard English and Belfast English is shared, it makes little sense to propose two complete linguistic systems in competition in Belfast. Rather, varieties may differ from each other in small ways, allowing for different settings of "micro-parameters" such as the mechanism of subject-verb agreement. Henry makes use of an early version of the Minimalist Program (Chomsky 1993) in which syntactic operations are motivated by the need for lexical constituents (nouns and verbs) to move to functional positions (tense, agreement) with matching features. In her analysis, whether the verb moves to a higher functional position where it can check that feature to achieve agreement depends on the strength of that feature. Thus, the variation in subject-verb agreement in (9.1) results from optionally strong or weak settings for the feature on the functional category. Although the results of mini-grammars and optional microparametric variation are empirically the same, they differ in terms of the degree to which optionality is admitted as part of the linguistic system.

Many linguists resist allowing <u>any</u> optionality in a theory of language. It is not clear whether this resistance is an assumption or a necessary part of linguistic theory-building (David Adger, personal communication), but Embick (2008) argues that several principles of generativist theory result in the property of Single Output (SO), defined in (9.2).

(9.2) Single Output
An input N_I to a derivation C_{HL} yields a single output N_O.
(adapted from Embick 2008: 65)

Single Output states that, given some input structure, the computational system (C_{HL}) can produce one and only one output. The implication of Single Output is that there is no optionality in the computational system of language and that, if there are two outputs derived from the same input, there must be more than one linguistic system present. A further implication of Single Output is that, since only one output can eventually be uttered, the mechanism for choosing among outputs (i.e. variants) must lie outside of language. Most approaches to accommodating variation in linguistic theory that have been proposed, while accepting some form of Single Output, take advantage of theory-internal mechanisms to derive the variation. In the following sections, we examine some of these approaches in the domains of phonology and grammar.

9.1.1 Variation in Phonological Theory

In phonological theory, there have been a number of approaches to model variation by working within the framework of Optimality Theory (OT; Prince & Smolensky 2004). OT conceptualizes linguistic processes as the outcome of ranked constraints, each of which can be violated to satisfy a higher-ranked constraint. For example, Nagy and Reynolds (1997) model variable word-final deletion in Faetar, a Francoprovençal dialect spoken in Italy. A word like /bró.kə.lə/ "fork" is variably pronounced as [brókələ], [brókəl], [brókə], or [brok]. Nagy and Reynolds propose that the different realizations result from different orderings of the constraints on representation, with Align-PrWd, defined in (9.3), as the crucial constraint differentiating realizations.

(9.3) Align-PrWd
The right boundary of a prosodic word coincides with the right edge of the head or main-stressed syllable.
(Nagy & Reynolds 1997: 42)

The form [brok] satisfies Align-PrWd (since the main stressed syllable occurs at the right edge of the word), but it violates another constraint, Parse, which says that all segmental material present in the input must appear in the output. In contrast, the form [brókələ] satisfies Parse but violates Align-PrWd (since two syllables intervene between the stressed syllable and the right edge of the word). Thus, the acceptability of each form depends on the relative ranking of constraints. Nagy and Reynolds suggest that the variability arises because Align-PrWd can "float" throughout the ranking, yielding multiple possible rankings and therefore multiple possible outputs. Since each ranking represents a different "grammar" in OT, floating constraints essentially achieves the same effects as the multiple (mini-)grammars discussed above. Table 9.1 shows

Table 9.1 Observed and expected rates of variation for different forms of "fork" in Faetar (Nagy & Reynolds 1997: 44).

Output:	Expected %	Observed %	Difference
[brókələ]	57	55	2
[brókəl]	11	15	4
[brókə]	11	14	3
[brok]	21	10	11

a comparison of the observed and expected relative frequencies of forms for this lexical item.

Note that the prediction for [brókələ] is surprisingly close, with only a 2 percent difference between expected and observed frequencies, and the expected frequencies for [brókəl] and [brókə] are similar to the observed frequencies. However, the quantitative predictions for [brok] do not match the observed variation particularly well. Another problem with the "floating constraints" approach is that it seems to assume that each ranking has the same probability of occurring. In a sense, all constraints are weighted equally in their contribution to the output, suggesting that each ranking should occur randomly.

Alternative approaches to variation in OT not only rank constraints with respect to each other but also provide each constraint with a numerical weighting (either positive or negative). Coetzee and Pater (in press) survey a number of such approaches within OT. Stochastic OT (Boersma 1997), Noisy Harmonic Grammar (Boersma & Weenink 2008), and Maximum Entropy Harmonic Grammar (Johnson 2002) use slightly different mechanisms for calculating numeric weighting (see Coetzee & Pater, in press, for a detailed discussion), but they all allow learners to derive weightings on the basis of repeated exposure to output forms. Using a computer-based learning algorithm, they test the predictions of these three approaches on the relative effects of the following phonological context on (t/d)-deletion, based on the constraints in (9.4).

(9.4) a. *Ct — Assign a violation mark to a consonant cluster ending in a coronal stop.
b. Max — Assign a violation mark to an input consonant that is not present in the output.
c. Max-Pre-V — Assign a violation mark to an input consonant in pre-vocalic position that is not present in the output.
d. Max-Final — Assign a violation mark to an input consonant in phrase final position that is not present in the output.

Table 9.2 Observed and predicted rates of (t/d)-deletion in two dialects of English according to different implementations of Optimality Theory (OT) learning algorithms: Stochastic OT (St-OT), Noisy Harmonic Grammar (N-HG) and Maximum Entropy Harmonic Grammar (ME-HG) (adapted from Coetzee & Pater, in press).

		*Ct	Max-P-V	Max-Fin	Max	**Following Phonological Context**		
						Vowel	Pause	Consonant
New York	*Observed:*					*.66*	*.83*	*1.00*
	St-OT	107.6	106.5	104.9	92.4	.66	.84	1.00
	N-HG	141.1	80.9	79.0	58.9	.65	.83	1.00
	ME-HG	140.4	80.3	79.3	59.6	.65	.83	1.00
Philadelphia	*Observed:*					*.38*	*.12*	*1.00*
	St-OT	107.2	108.2	110.6	92.8	.37	.12	1.00
	N-HG	139.2	79.4	82.4	60.8	.38	.12	1.00
	ME-HG	139.5	79.5	81.0	60.5	.38	.12	1.00

Based on Guy's (1980) finding of different effects of following pause on (t/d)-deletion in Philadelphia and New York, they differentiate the rankings by dialect. Table 9.2 (adapted from Coetzee and Pater's table) reports their comparison of the three OT models for the relative ranking of constraints for the following phonological context in these two dialects. As the final right-hand columns show, all three models make predictions that closely match the observed distributions in each dialect. Thus, approaches to variation within OT that provide numeric weightings of constraints provide a better fit to the data than the floating-constraint approach. However, although they account for the effect of one factor group (following phonological context), the contextual factors must be encoded within the theory as constraints (Max-Pre-V and Max-Final). It remains to be seen whether the predictions of these approaches are as successful when multiple factors are taken into account simultaneously.

9.1.2 Variation in Grammatical Theory

Approaches to accommodating variation in grammatical theory have tended to maintain Single Output, resorting to other theory-internal mechanisms to derive the variation. In contrast to Optimality Theory, which allows the computational system (constraint ranking) to be changed in order to produce different outputs, the Minimalist Program (Chomsky 1995) assumes that once the lexical and functional elements of a sentence have been selected and merged into a syntactic representation, only one output is possible. Minimalist approaches to variation therefore assume that variation can arise only at two points in the derivation: the (phonological and morphological) spell-out of functional features, or the selection of elements (features or words) from the lexicon.

Consider the variation in the paradigm of past-tense *be* in the English of Buckie, Scotland (Adger & Smith 2005), shown in Table 9.3, which has been used in recent discussions of grammatical variation in linguistic theory. This paradigm exhibits categorical *was* in two contexts (first person singular and third person singular), categorical *were* in one context (third person plural pronoun subject) and variable *was/were* elsewhere. Any theoretical model of this paradigm must therefore be able to account for both the categorical and variable facts.

In an approach based on Distributed Morphology (Halle & Marantz 1993), Nevins and Parrott (2009) argue that this paradigm can be described through the interplay of three features, [±Author], [±Participant], [±Plural], each combination of which receives a particular phonological realization captured in the spell-out rules shown in (9.5). According to these rules, if a feature combination of [+Participant, –Author] or [+Plural] exists in the syntactic tree for the verb *be*, pronounce it as *were*; otherwise, pronounce it as *was*.

(9.5)		Features		Spell-out
	a.	[+Participant, –Author]	⇔	*were*
	b.	[+Plural]	⇔	*were*
	c.	*elsewhere*	⇔	*was*

In order to derive the variability, they propose an additional "impoverishment rule" that (variably) removes the person and number features when a [+Participant] feature is present. As a result of this rule, a form with a [+Participant] feature, which is normally the input for rule (9.5a), instead variably becomes the input to rule (9.5c), and may be pronounced as either *was* or *were*. This approach allows us to derive variation without having to violate Single Output, but it does have theoretical and empirical shortcomings. On a theoretical level, it seems arbitrary to propose a feature that (variably) strips syntactic representations of features. There does not appear to be any mechanism restricting the types of impoverishment rules. On an empirical level, the impoverishment rule

Table 9.3 Paradigm of *was/were* agreement in Buckie English (Adger & Smith 2005).

	Singular	**Plural**
1st	I was	we was/were
2nd	you was/were	you was/were
3rd pronoun	(s)he was	they were

seems to operate randomly, suggesting that the variation should occur at equal rates in all person-number-subject contexts (see the discussion of “free variation” in Chapter 3). But as Table 9.4 shows, there are very different rates of *was* in each grammatical person. Thus, although this approach allows for multiple forms to express the same meaning, it does not make any reliable quantitative predictions about the distribution of forms.

Working with the same data but within a later version of the Minimalist Program (Chomsky 1995), Adger (2006) argues that we need nothing beyond what is already available in the theory to derive the variation. He uses a set of features similar to those of Nevins and Parrott (2009) along with a spellout rule for each:

(9.6)		Feature	Spell-out
	a.	[singular:+]	*was*
	b.	[singular:–]	*were*
	b.	[participant:+]	*was*
	c.	[author:–]	*were*
	g.	[author:+]	*was*

Since there is a great deal of homophony in this system, a form may be spelled out based on a number of different feature combinations. For example, as shown in (9.7), second person singular *you* and first person plural *we* are twice as likely to be pronounced as *was* than as *were* (9.7a, 9.7c), while second person plural *you* is twice as likely to be pronounced as *were* than as *was*. If we compare these expected frequencies with the observed frequencies, as in Table 9.5, we see a fairly close match for *you* (singular) and *we*, though not for *you* (plural).

Table 9.4 Distribution of *was* in *were* contexts by grammatical person (Adger & Smith 2005: 156).

	%	N
2nd singular *you*	69	161
1st plural *we*	67	368
3rd plural *they*	0	435
Existential *there*	90	162
NP plural	56	187

Table 9.5 Expected and observed distribution of *was* in Buckie English (Adger 2006: 522).

Pronoun	Expected %	Observed %	N
2nd singular	67	69	161
1st plural	67	67	368
2nd plural	33	10	10

(9.7)		Features	Spell-out
a.	*you* (sg.)	[singular:+, participant:+, author:–]	*was*, *was*, *were*
b.	*you* (pl.)	[singular:–, participant:+, author:–]	*were*, *was*, *were*
c.	*we*	[singular:–, participant:+, author:+]	*were*, *was*, *was*

Adger's approach is promising, in that it requires no additional theoretical machinery to work, and it is also able to make predictions about the quantitative distribution, although those predictions do not always match the observed distribution closely. However, it is not clear what governs the choice of features from the lexicon in the first place. Is it completely random or are there other linguistic or extralinguistic constraints?

9.1.3 Variation and Linguistic Theory

The interface between linguistic theory and linguistic variation is a promising and exciting field of future research, one that requires expertise from various subfields in order to achieve both descriptive and explanatory success. From the perspective of variationist research, there are a number of considerations that need to be taken into account in any linguistic theory that tries to accommodate variation.

First, where is variation located? Is it within the linguistic system, or is there an outside mechanism? Is lack of optionality (Single Output) a necessary part of linguistic theory? Absent a principled method for distinguishing code-switching or grammar competition from true variation, we must consider seriously the possibility of optionality in the linguistic system.

Second, what are the relative roles of formal and functional constraints? We have seen evidence that functional constraints may influence the variation, but that should not lead us to abandon the notion of linguistic structure altogether. At the same time, we cannot view the linguistic system as operating entirely without reference to functional considerations. A possible compromise is to adopt the "soft modularity" proposed in Torres Cacoullos and Walker (2009a), in which there are different modules of the grammar, but they may interact with each other in various (non-discrete) ways.

Finally, we want to be able to make quantitative predictions. While we do not expect an exact numerical match of the expected and observed distributions, we do not want them to be significantly different. One of the drawbacks of the theoretical models reported here is that they tend to be statistically quite simple, relying on a single factor group, without any tests of statistical significance or goodness of fit. As we saw in Chapter 4, multiple factors may operate simultaneously to affect the variation. Any theory of language that tries to accommodate variation must therefore avail itself of the methodological and statistical tools of variationist analysis. In the end, any predictions made by a theory must be empirically verifiable.

9.2 Conclusion

In this chapter, we have considered whether the variationist method can be accommodated within linguistic theory. Although it is tempting to rely solely on functional constraints on language, we cannot entirely do away with formal or structural considerations. Linguistic theory tends to view variation as a problem, often dealt with by relegating it to outside the linguistic system or arguing for the co-existence of multiple linguistic systems. Since many linguists resist optionality in linguistic systems, attempts to model variation within linguistic theory tend to rely on theory-internal mechanisms such as constraint ranking or feature specification. Although linking linguistic theory and linguistic variation provides a promising area of future research, any theoretical account of linguistic variation must be empirically verifiable to successfully bridge linguistic theory and variationist linguistics.

9.3 Further Reading

Adger, David. 2006. Combinatorial variability. *Journal of Linguistics* 42: 503–30.

Embick, David. 2008. Variation and morphosyntactic theory: Competition fractionated. *Language and Linguistics Compass* 2/1: 59–78.

Gardner-Chloros, Penelope. 2009. *Code Switching*. Cambridge: Cambridge University Press.

Henry, Alison. 1995. *Belfast English and Standard English: Dialect Variation and Parameter Setting*. New York and Oxford: Oxford University Press.

Nagy, Naomi and Bill Reynolds. 1997. Optimality Theory and variable word-final deletion in Faetar. *Language Variation and Change* 9: 37–55.

Nevins, Andrew and Jeffrey K. Parrott. 2009. Variable rules meet impoverishment theory. *Lingua* 119.

Notes

2 Variation and Variables

1 Thanks to Ian Smith for providing this example.

2 Thanks to Abdel-Wahhab Zraouti for providing this example from Moroccan Arabic.

3 Let us assume that /z/ is the underlying form in order to make the formulation of the rules less complicated.

4 Note that this definition of the variable context yields a different set of data from the form-based definition in the previous paragraph. As we will see in Chapter 3, since we now have at least four variants, the overall relative frequency of each form will be different.

3 The Analysis of Linguistic Variation

1 Current linguistic theory has largely abandoned rules, or has limited rules to a relatively small and simple set (e.g. “move something”), and now views linguistic processes as the outcome of ranked constraints on structure (such as Optimality Theory (McCarthy & Prince 1993; Prince & Smolensky 2004) or the requirement of lexically-specified features to occur at particular points in the sentence (the Minimalist Program (Chomsky 1995)). We discuss the application of these models to linguistic variation in Chapter 9.

4 Multivariate Analysis with GoldVarb

1 Note that the same code cannot be used twice within the same factor group, but the same code can be used in different factor groups.

2 Conventions for transcribing utterances vary from researcher to researcher. As you can see from the fragment in Figure 4.1, I represent the phonetic realization of the token and highlight the word in which it occurs using all capital letters.

3 GoldVarb requires that you choose one of the factors in each factor group as a “default” value, but it does not matter which you choose.

4 Here and in the rest of this chapter, we use the Macintosh version of GoldVarb X. The PC/Windows version of GoldVarb X has the same functions as the Macintosh version, but the implementation of the functions differs.

5 Trinomial analysis, involving three results, is available in some of the other implementations of VARBRUL.

6 Another indication of the relative strength of a factor group in the analysis is the order of selection in the step-up procedure. The stronger the effect, the earlier the factor group is selected.

7 In the Macintosh implementation of GoldVarb, the cross-tabulation function is located in the Cells pulldown menu.

8 We can test whether this difference is statistically significant by using twice the difference between the two log likelihoods as a chi-square value. The degrees of freedom is the difference between the degrees of freedom for each run, which itself is derived by subtracting the total number of factor groups from the total number of factors (see Guy 1993 and Paolillo 2002 for a more thorough discussion).

6 Variation in Grammatical Systems

1 These example sentences illustrate what Tagliamonte (2006a: 96) calls "super tokens": "alternation of variants by the same speaker in the same stretch of discourse."

2 In this table, and in other tables in this book, I follow the convention of indicating with empty brackets those factor groups not selected as significant.

8 Language Contact

1 Meyerhoff and Walker (2007) also provide evidence that the individual speakers within each community behave similarly.

References

Adank, Patti, Roel Smits and Roeland van Hout. 2004. A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America* 116: 3099–3107.

Adger, David. 2006. Combinatorial variability. *Journal of Linguistics* 42: 503–30.

Adger, David and Jennifer Smith. 2005. Variation and the Minimalist Program. In Leonie Cornips and Karen Corrigan (eds.), *Syntax and Variation: Reconciling the Biological and the Social*. Amsterdam: Benjamins, 149–78.

Angermeyer, Philipp and John V. Singler. 2003. The case for politeness: Pronoun variation in co-ordinate NPs in object position in English. *Language Variation and Change* 15: 171–209.

Appel, René and Pieter Muysken. 1987. *Language Contact and Bilingualism*. New York: Edward Arnold.

Baayen, Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Bailey, Charles-James N. 1973. *Variation and Linguistic Theory*. Arlington, VA: Center for Applied Linguistics.

Bailey, Guy, Natalie Maynor and Patricia Cukor-Avila (eds.). 1991. *The Emergence of Black English: Texts and Commentary*. Philadelphia/Amsterdam: Benjamins.

Bayley, Robert. 1996. Competing constraints on variation in the speech of adult Chinese learners of English. In Robert Bayley and Dennis R. Preston (eds.), *Second Language Acquisition and Linguistic Variation*. Amsterdam/Philadelphia: Benjamins, 97–120.

Bayley, Robert. 2002. The quantitative paradigm. In J.K. Chambers, Peter Trudgill and Natalie Schilling–Estes (eds.), *The Handbook of Language Variation and Change*. Oxford: Blackwell, 117–41.

Bickerton, Derek. 1971. Inherent variability and variable rules. *Foundations of Language* 7: 457–92.

Bickerton, Derek. 1975. *Dynamics of a Creole System*. Cambridge: Cambridge University Press.

Bickerton, Derek. 1981. *Roots of Language*. Ann Arbor, MI: Karoma.

Bickerton, Derek. 1984. The language bioprogram hypothesis. *The Behavioral and Brain Sciences* 7: 173–221.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 21: 43–58.

Boersma, Paul and David Weenink. 2008. *Praat: Doing Phonetics by Computer*. http://www.praat.org (accessed November 11, 2009)

Bolinger, Dwight. 1977. *Meaning and Form*. London/New York: Longman.

Bybee, Joan. 2000. The phonology of the lexicon: Evidence from lexical diffusion. In Michael Barlow and Suzanne Kemmer (eds.), *Usage-Based Models of Language*. Stanford, CA: CSLI Publications, 65–85.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82: 711–33.

Bybee, Joan, Revere Perkins and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Cameron, Richard. 1993. Ambiguous agreement, functional compensation, and nonspecific *tú* in the Spanish of San Juan, Puerto Rico, and Madrid, Spain. *Language Variation and Change* 5: 305–334.

Cedergren, Henrietta. 1972. *The Interplay of Social and Linguistic Factors in Panama*. Ph.D. Dissertation, Cornell University.

Cedergren, Henrietta and David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50: 333–55.

Chambers, J.K. 1973. Canadian Raising. *Canadian Journal of Linguistics* 18: 113–35.

Chambers, J.K. 2008. *Sociolinguistic Theory*. Revised edition. Oxford/Malden, MA: Blackwell.

Chambers, J.K., Peter Trudgill and Natalie Schilling-Estes (eds.). 2002. *The Handbook of Language Variation and Change*. Oxford: Blackwell.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1993. A minimalist program for linguistic theory. In K. Hale and S.J. Keyser (eds.), *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*. Cambridge, MA: MIT Press.

Clarke, Sandra, Ford Elms and Amani Youssef. 1995. The third dialect of English: Some Canadian evidence. *Language Variation and Change* 7: 209–28.

Clermont, Jean and Henrietta Cedergren. 1979. Les "R" de ma mere sont perdus dans l'air. In P. Thibault (ed), *Le français parlé: Études sociolinguistiques*. Edmonton, Alberta: Linguistic Research, 13–28.

Coetzee, Andries and Joe Pater. in press. The place of variation in phonological theory. In John Goldsmith, Jason Riggle and Alan Yu (eds.), *The Handbook of Phonological Theory, Second Edition*. Oxford: Blackwell.

Coulmas, Florian. 1997. *The Handbook of Sociolinguistics*. Oxford/Malden, MA: Blackwell.

D'Arcy, Alexandra. 2005. *Like: Syntax and Development*. Ph.D. dissertation, University of Toronto.

DeCamp, David. 1971. Toward a generative analysis of a post-creole speech continuum. In Dell Hymes (ed.), *Pidginization and Creolization of Languages*. Cambridge: Cambridge University Press, 349–370.

Eckert, Penelope. 1999. *Linguistic Variation as Social Practice*. Oxford: Blackwell.

Eckert, Penelope and John R. Rickford (eds.). 2001. *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Ellegård, Alvar. 1953. *The Auxiliary* do: *The Establishment and Regulation of its Use in English*. Gothenburg Studies in English, 2. Stockholm: Almqvist & Wiksell.

Embick, David. 2008. Variation and morphosyntactic theory: Competition fractionated. *Language and Linguistics Compass* 2/1: 59–78.

Francis, W. Nelson and Henry Kučera. 1982. *Frequency Analysis of English Usage*. Boston: Houghton Mifflin.

Gardner-Chloros, Penelope. 2009. *Code Switching*. Cambridge: Cambridge University Press.

Gumperz, John J. and Robert Wilson. 1971. Convergence and creolization: A case study from the Aryan/Dravidian border in India. In Dell Hymes (ed.), *Pidginization and Creolization of Languages*. Cambridge: Cambridge University Press, 151–67.

Guy, Gregory R. 1980. Variation in the group and the individual: The case of final stop deletion. In William Labov (ed.), *Locating Language in Time and Space*. New York: Academic Press, 1–36.

Guy, Gregory R. 1988. Advanced VARBRUL analysis. In K. Ferrara, B. Brown, K. Walters and J. Baugh (eds.), *Linguistic Change and Contact*. Austin: Department of Linguistics, University of Texas at Austin, 124–36.

Guy, Gregory R. 1991. Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change* 3:1–22.

Guy, Gregory R. 1993. The quantitative analysis of linguistic variation. In D. Preston (ed.), *American Dialect Research*. Amsterdam: Benjamins, 223–49.

Guy, Gregory R. and Charles Boberg. 1997. Inherent variability and the obligatory contour principle. *Language Variation and Change* 9: 149–64.

Guy, Gregory R., Crawford Feagin, Deborah Schiffrin and John Baugh (eds.). 1996. *Towards a Social Science of Language, Volume 1: Variation and Change in Language and Society*, Amsterdam/Philadelphia: Benjamins.

Halle, Morris and Alec Marantz. 1993. Distributed Morphology and the pieces of inflection. In Kenneth Hale and Samuel Jay Keyser (eds.), *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. Cambridge, MA: MIT Press, 111–76.

Harvie, Dawn. 1998. Null subject in English: Wonder if it exists? *Cahiers linguistiques d'Ottawa* 26: 15–25.

Henry, Alison. 1995. *Belfast English and Standard English: Dialect Variation and Parameter Setting*. New York/Oxford: Oxford University Press.

Hoffman, Michol F. in preparation. The progress of the Canadian Shift in Toronto. Ms., York University.

Hoffman, Michol F. and James A. Walker. in press. Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. *Language Variation and Change*.

Holm, John A. 1988. *Pidgins and Creoles, Volume 1: Theory and Structure*. Cambridge: Cambridge University Press.

Holm, John et al. 2000. The creole verb: A comparative study of stativity and time reference. In John McWhorter (ed.), *Language Change and Language Contact in Pidgins and Creoles*. Amsterdam: Benjamins, 133–62.

Hopper, Paul J. 1991. On some principles of grammaticalization. In Elisabeth C.

Traugott and Bernd Heine (eds.), *Approaches to grammaticalization*. Amsterdam: John Benjamins, 17–35.

Hopper, Paul J. 1998. Emergent Grammar. In Michael Tomasello (ed.), *The New Psychology of Language*. Mahwah, NJ: Lawrence Erlbaum Associates, 155–75.

Hopper, Paul J. and Elizabeth C. Traugott. 2003. *Grammaticalization*. Cambridge: Cambridge University Press.

Inkelas, Sharon and Draga Zec. 1993. Auxiliary reduction without empty categories: A prosodic account. *Working Papers of the Cornell Phonetics Laboratory* 8:205–253.

Johnson, Mark. 2002. Optimality-theoretic Lexical Functional Grammar. In Suzanne Stevenson and Paolo Merlo (eds.), *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*. Amsterdam: John Benjamins, 59–73.

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1: 199–244.

Kroch, Anthony. 2000. Syntactic change. In Mark Baltin (ed.), *The Handbook of Contemporary Syntactic Theory*. Malden, MA/Oxford: Blackwell, 629–739.

Kroch, Anthony. 2001. Syntactic change. In Mark Baltin and Christopher Collins (eds.), *The handbook of* contemporary syntactic theory. Oxford/Malden, MA: Blackwell, 699–729

Labov, William. 1963. The social motivation of a sound change. *Word* 19: 273–309.

Labov, William. 1966. *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45: 715–62.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. 1984. Field methods of the project on linguistic change and variation. In John Baugh and Joel Sherzer (eds.), *Language in Use: Readings in Sociolinguistics*. Englewood Cliffs, NJ: Prentice Hall, 28–53.

Labov, William. 1989. The child as linguistic historian. *Language Variation and Change* 1: 85–97.

Labov, William. 1994. *Principles of Linguistic Change, Volume 1: Internal Factors*. Oxford: Blackwell.

Labov, William. 1998. Co-existent systems in African-American vernacular English. In Salikoko S. Mufwene et al. (eds.), *African-American Vernacular English*. London: Routledge, 110–53.

Labov, William. 2008. Plotnik. http://www.ling.upenn.edu/~wlabov (accessed November 11, 2009).

Labov, William, P. Cohen, C. Robins and J. Lewis. 1968. *A Study of the Nonstandard English of Negro and Puerto Rican Speakers in New York City*. Cooperative Research Report 3288, Vol I. Philadelphia: U.S. Regional Survey.

Labov, William, Malcah Yaeger and Richard Steiner. 1972. *A Quantitative Study of Sound Change in Progress* (Vol. 1). Philadelphia: U.S. Regional Survey.

Laks, Bertrand. 1992. La linguistique variationniste comme méthode. *Langages* 108: 34–50.

Lavandera, Beatriz. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7: 171–82.

Levey, Stephen. 2006. Visiting London relatives. *English World-Wide* 27: 45–70.

Lightfoot, David. 1979. *Principles of Diachronic Syntax*. Cambridge: Cambridge University Press.

Lightfoot, David. 1991. *How to Set Parameters: Arguments from Language Change*. Cambridge, MA: MIT Press.

McCarthy, John J. and Andrew Prince. 1993. Prosodic morphology I: Constraint interaction and satisfaction. Unpublished ms., University of Massachusetts, Amherst, and Rutgers University.

Meillet, Antoine. 1912. L'evolution des formes grammaticales. *Rivista di scienza* XII 26/6. Reprinted in 1948. *Linguistique historique et linguistique générale*. Paris: Champion, 130–48.

Meyerhoff, Miriam. 2000. The emergence of creole subject-verb agreement and the licensing of null subjects. *Language Variation and Change* 12: 203–30.

Meyerhoff, Miriam, Jack Sidnell and James A. Walker. in preparation. Varieties of English on Bequia, St. Vincent and the Grenadines. Ms., University of Edinburgh / University of Toronto / York University.

Meyerhoff, Miriam and James A. Walker. 2007. The persistence of variation in individual grammars: Copula absence in "urban sojourners" and their stay-at-home peers, Bequia (St. Vincent and the Grenadines). *Journal of Sociolinguistics* 11: 346–66.

Milroy, Lesley and Matthew J. Gordon. 2003. *Sociolinguistics: Method and Interpretation*. Oxford/Malden, MA: Blackwell.

Muysken, Pieter. 2002. *Bilingual speech: A typology of code-mixing*. Cambridge: Cambridge University Press.

Nagy, Naomi and Bill Reynolds. 1997. Optimality Theory and variable word-final deletion in Faetar. *Language Variation and Change* 9: 37–55.

Naro, Anthony. 1981. The social and structural dimensions of a syntactic change. *Language* 57: 63–98.

Naro, Anthony, Edair Görski and Eulália Fernandes. 1999. Change without change. *Language Variation and Change* 11: 197–211.

Nespor, Marine and Irene Vogel. 1986. *Prosodic Phonology*. Dordrecht: Foris.

Nevins, Andrew and Jeffrey K. Parrott. 2009. Variable rules meet impoverishment theory. *Lingua* 119.

Noble, Shawn. 1985. To have and have got. Paper presented at New Ways of Analyzing Variation 14, Georgetown University, Washington, DC.

Oliveira e Silva, Giselle. 1982. *Estudo da regularidade na variação dos possessives no português do Rio de Janeiro*. Ph.D. Dissertation, Universidade Federal do Rio de Janeiro.

Paolillo, John C. 2002. *Analyzing Linguistic Variation: Statistical Models and Methods*. Stanford: CSLI Publications.

Pappas, Panayiotis. 2008. Object clitic placement in Cypriot Greek: Results from a variationist analysis. Paper presented at New Ways of Analyzing Variation 38, Rice University, Houston, TX.

Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In Joan Bybee and Paul Hopper (eds.), *Frequency Effects and the Emergence of Lexical Structure*. Amsterdam: Benjamins, 137–57.

Poplack, Shana. 1980a. Deletion and disambiguation in Puerto Rican Spanish. *Language* 56: 371–85.

Poplack, Shana. 1980b. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a typology of code-switching. *Linguistics* 18: 581–618.

Poplack, Shana. 1989. The care and handling of a megacorpus: The Ottawa-Hull French Project. In Ralph Fasold and Deborah Schiffrin (eds.), *Language Change and Variation*. Amsterdam/Philadelphia: Benjamins, 411–44.

Poplack, Shana. 1992. The inherent variability of the French subjunctive. In Christiane Laeufer and Terrell A. Morgan (eds.), *Theoretical Analyses in Romance Linguistics*. Amsterdam: John Benjamins, 235–63.

Poplack, Shana. 1993. Variation theory and language contact: concepts, methods, and data. In Dennis R. Preston (ed.), *American Dialect Research*. Amsterdam: Benjamins, 251–86.

Poplack, Shana. 1997. The sociolinguistic dynamics of apparent convergence. In Gregory R. Guy et al. (eds.), *Towards a Social Science of Language: Papers in Honor of William Labov*. Amsterdam/Philadelphia: Benjamins, 285–309.

Poplack, Shana (ed.). 2000. *The English History of African American English*. Oxford/Malden, MA: Blackwell.

Poplack, Shana and Elisabete Malvar. 2007. Elucidating the transition period in linguistic change: The expression of the future in Brazilian Portuguese. *Probus* 19:121–69.

Poplack, Shana and Marjory Meechan. 1998. Introduction: How languages fit together in codemixing. *International Journal of Bilingualism* 2:127–38.

Poplack, Shana and David Sankoff. 1987. The Philadelphia story in the Spanish Caribbean. *American Speech* 64: 291–314.

Poplack, Shana, David Sankoff and Christopher Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics* 26:47–104.

Poplack, Shana and Sali Tagliamonte. 1991. African American English in the diaspora: The case of old-line Nova Scotians. *Language Variation and Change* 3: 301–39.

Poplack, Shana and Sali Tagliamonte. 1996. Nothing in context: variation, grammaticalization and past time marking in Nigerian Pidgin English. In Philip Baker (ed.) *Changing Meanings, Changing Functions. Papers relating to grammaticalization in contact languages*. Westminster, UK: University Press, 71–94.

Poplack, Shana and Sali Tagliamonte. 1999. The grammaticization of *going to* in (African American) English. *Language Variation and Change* 11: 315–42.

Poplack, Shana and Sali Tagliamonte. 2001. *African American English in the Diaspora*. Oxford and Malden, MA: Blackwell.

Poplack, Shana and Danielle Turpin. 1999. Does the FUTUR have a future in (Canadian) French? *Probus* 11.133–64.

Poplack, Shana, James A. Walker and Rebecca Malcolmson. 2006. An English "like no other"? Language contact and change in Quebec. *Canadian Journal of Linguistics* 51: 185–213

Preston, Dennis R. 1989. *Sociolinguistics and Second Language Acquisition*. Oxford: Blackwell.

Preston, Dennis R. (ed.). 1993. *American Dialect Research*. Amsterdam/Philadelphia: Benjamins.

Prince, Alan and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Malden, MA/Oxford: Blackwell.

Richardson, Carmen. 1991. Habitual structures among Blacks and Whites in the 1990s. *American Speech* 66: 292–302.

Rickford, John R. 1985. Ethnicity as a sociolinguistic boundary. *American Speech* 60: 99–125.

Romaine, Suzanne. 1984. On the problem of syntactic variation and pragmatic meaning in sociolinguistic theory. *Folia Linguistica* 18: 409–39.

Sankoff, David. 1988a. Sociolinguistics and syntactic variation. In F.J. Newmeyer (ed.) *Linguistics: The Cambridge Survey. Volume IV, Language: The Sociocultural Context*. Cambridge: Cambridge University Press, 140–61.

Sankoff, David. 1988b. Variable rules. In Ulrich Ammon, Norbert Dittmar and Klaus J. Mattheier (eds.), *Sociolinguistics: An International Handbook of the Science of Language and Society*. Berlin/New York: Walter de Gruyter, 984–97.

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

Sankoff, David and Pierrette Thibault. 1981. Weak complementarity: Tense and aspect in Montreal French. In B. Strong Johns and D. Strong (eds.), *Syntactic Change*. Ann Arbor: University of Michigan, 206–16.

Sankoff, David, Sali Tagliamonte and Eric Smith. 2005. *GoldVarb X: A Multivariate Analysis Application*. Department of Linguistics, University of Toronto / Department of Mathematics, University of Ottawa.

Sankoff, Gillian. 1974. A quantitative paradigm for the study of communicative competence. In Richard Bauman and Joel Sherzer (ed.) *Explorations in the Ethnography of Speaking*. Cambridge: Cambridge University Press, 18–49.

Sankoff, Gillian. 1980. *The Social Life of Language*. Ann Arbor, MI: Karoma.

Sankoff, Gillian. 1990. The grammaticalization of tense and aspect in Tok Pisin and Sranan. *Language Variation and Change* 2: 295–312.

Sankoff, Gillian. 2006. Age: Apparent time and real time. In Keith Brown (ed.) *The Encyclopedia of Languages and Linguistics, Vol. 1*. Cambridge: Elsevier, 110–16.

Sankoff, Gillian and Hélène Blondeau. 2007. Language change across the lifespan: /r/ in Montreal French. *Language* 83:560–88.

Sankoff, Gillian and Pierrette Thibault. 1977. L'alternance entre les auxiliaires avoir et être en français parlé à Montréal. *Langue française* 34: 81–108.

Santa Ana, Otto. 1991. *Phonetic Simplification Processes in English of the Barrio: A Cross-Generational Sociolinguistic Study of the Chicanos of Los Angeles*. Ph.D. Dissertation, University of Pennsylvania.

Sapir, Edward. 1921. *Language: An Introduction to the Study of Speech*. New York: Harcourt, Brace and Company.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Selinker, Larry. 1972. Interlanguage. *International Review of Applied Linguistics* 10: 219–31.

Selkirk, Elisabeth. 1984. *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, MA: MIT Press.

Silva, David. 1994. The variable elision of unstressed vowels in European Portuguese: A case study. *UTA Working Papers in Linguistics* 1:79–94.

Tagliamonte, Sali. 2006a. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Tagliamonte, Sali. 2006b. “So cool, right?”: Canadian English entering the 21st century. *Canadian Journal of Linguistics* 51(2/3): 309–31.

Tagliamonte, Sali and Alexandra D’Arcy. 2007. Frequency and variation in the community grammar: Tracking a new change through the generations. *Language Variation and Change* 19: 199–217.

Tagliamonte, Sali and Rachel Hudson. 1999. *Be like* et al. beyond America: The quotative system in British and Canadian youth. *Journal of Sociolinguistics* 3: 147–72.

Thomason, Sarah A. and Terence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley, CA: University of California Press.

Torres Cacoullos, Rena. in press. Variation and grammaticalization. In Manuel Diaz-Campos (ed.), *The handbook of Hispanic sociolinguistics*. Oxford/ Malden, MA: Blackwell.

Torres Cacoullos, Rena and James A. Walker. 2009a. On the persistence of grammar in discourse formulas: A variationist study of *that*. *Linguistics* 47: 1–43.

Torres Cacoullos, Rena and James A. Walker. 2009b. The present of the English future: Grammatical variation collocations in discourse. *Language* 85.

Trudgill, Peter. 2000. *Sociolinguistics: An Introduction to Language and Society*. London: Penguin.

Van Herk, Gerard and James A. Walker. 2005. S marks the spot? Regional variation and early African American correspondence. *Language Variation and Change* 17(2): 113–31.

Vincent, Diane and David Sankoff. 1992. Punctors: A pragmatic variable. *Language Variation and Change* 4: 205–16.

Walker, James A. 1995. The /r/-ful truth about African Nova Scotian English. Paper presented at New Ways of Analyzing Variation 24, University of Pennsylvania, Philadelphia, PA. http://www.yorku.ca/jamesw/rless.pdf (accessed November 19, 2009).

Walker, James A. 2000a. Rephrasing the copula: Contraction and zero in early African American English. In Shana Poplack (ed.), *The English History of African American English*. Oxford/Malden, MA: Blackwell, 35–72.

Walker, James A. 2000b. *Present Accounted For: Prosody and Aspect in Early African American English*. Ph.D. Dissertation, University of Ottawa.

Walker, James A. 2008. Form, function, and frequency in phonology: (t/d)-deletion in Toronto English. Presented at New Ways of Analyzing Variation 37, Rice University, Houston, TX.

Warner, Anthony. 2005. Why DO dove: Evidence for register variation in Early Modern English negatives. *Language Variation and Change* 17: 257–80.

Weiner, E. Judith and William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19: 29–58.

Weinreich, Uriel, William Labov and Marvin I. Herzog. 1968. Empirical foundations for a theory of language change. In Winfred Lehmann and Igor Malkiel (eds.), *Directions for Historical Linguistics*. Austin: University of Texas Press, 95–195.

Wells, John Christopher. 1982. *Accents of English 1: An Introduction*. Cambridge: Cambridge University Press.

Wolfram, Walt. 1993. Identifying and interpreting variables. In Dennis R. Preston (ed.), *American Dialect Research*. Amsterdam: Benjamins, 193–221.

Wolfram, Walt and Eric Thomas. 2002. *The Development of African American English*. Oxford/Malden, MA: Blackwell.

Woods, Anthony, Paul Fletcher and Arthur Hughes. 1986. *Statistics in Language Studies*. Cambridge: Cambridge University Press.

Young, Richard and Robert Bayley. 1996. VARBRUL analysis for second language acquisition research. In D. Preston (ed.), *Second Language Acquisition and Linguistic Variation*. Amsterdam/Philadelphia: Benjamins, 253–306.

Zilles, Ana M.S. 2005. The development of a new pronoun: The linguistic and social embedding of *a gente* in Brazilian Portuguese. *Language Variation and Change* 17: 19–53.