5 The linguistic variable

How do you find a *linguistic variable*? This chapter will discuss the key construct in the variationist paradigm – the linguistic variable. It will detail the definition of a linguistic variable, describe what it is, how to identify it and how to circumscribe it.

DEFINING THE LINGUISTIC VARIABLE

> The definition of a linguistic variable is the first and also the last step in the analysis of variation. It begins with the simple act of noticing a variation – that there are two alternative ways of saying the same thing. (Labov to appear)

The most fundamental construct in variation analysis is the 'linguistic variable'. The quote above is the most recent one I could find from Labov himself; turning back to the original definition of the linguistic variable you find something a little more complicated. In 1966, Labov (1966/1982: 49) says the linguistic variable must be 'high in frequency, have a certain immunity from conscious suppression . . . [be] integral units of larger structures, and . . . be easily quantified on a linear scale'. Furthermore, the linguistic variable was required to be 'highly stratified' and to have 'an asymmetric distribution over a wide range of age levels or other ordered strata of the society' (Labov 1972c: 8). In this chapter, I shall 'unpack' what all this means. At the outset, however, the most straightforward and simple definition of the linguistic variable is simply 'two or more ways of saying the same thing' (Labov 1972c, Sankoff 1980: 55).

At the level of phonology, the linguistic variable is relatively straightforward. The alternates may simply differ by an extra phonological feature or two, such as the classic (t,d) and (ing) variables of English. Variable (t,d) involves word-final consonant clusters. Sometimes the

cluster is realised; sometimes it is not, as in (1a). Variable (ing) involves word-final *-ing*. Sometimes it is realised as [ŋ]; sometimes as [n], as in (1b). Variable (t) involves the pronunciation of word-internal intervocalic [t]. Sometimes it is realised as [t], sometimes as [ɾ], as in (1c). In these cases, there is little contention of semantic equivalence, i.e. 'means the same thing', since the variant forms alternate within the same word.

(1)
a. I misse[**t**] the bus yesterday. vs I miss[**Ø**] the bus yesterday.
b. shoppi[**ŋ**] vs shoppi[**n**]
c. bu[**t**]r vs bu[**ɾ**]r

In morphosyntax, however, alternation of forms may involve variable inflections, alternate lexical items or elementary syntactic differences that arise in the course of sentence derivation, as in (2). Is the original definition of the linguistic variable as 'two ways of saying the same thing' viable?

(2)
a. go slow**Ø** vs go slow**ly**
b. the woman ***who*** . . . vs the woman ***that*** . . .
c. he ***isn't*** vs he***'s not***

The question becomes whether or not two different ways of saying the same thing ever happens in syntax and semantics. If it does, how is it to be recognised, interpreted and explained effectively? Crucial to these questions is the often difficult task of defining the context of meaning, which requires having some principled way of dealing with the problematic relationship between linguistic form and linguistic function. Indeed, one of the key preoccupations of variation analysis has been that different forms can have the same meaning. But how can this be? Shouldn't each form have a different meaning?

From the very beginning, linguistics and sociolinguistics have been opposed in their treatment of 'meaning':

> two different lexical items or structures can almost always have some usages or contexts in which they have different meanings, or functions, and it is even claimed by some that this difference, though it may be subtle, is always pertinent whenever one of the forms is used. (Sankoff 1988b: 153)

The first recognition of the form/function problem is found in Weiner and Labov (1983). They demonstrate that generalised active sentences, as in (3a), and agentless passives, as in (3b), are opposing choices of the same syntactic variable.

(3)
a. They broke into the liquor closet.
b. The liquor closet was broken into.

In order to include these two variants in one syntactic variable, the two forms must have the same referential meaning. Such a supposition calls into question the nature of equivalence.

This is where there has been heated debate in the field, which has, in turn, been responsible for an evolution in thinking about variables. Much of this development occurred when analysts started studying linguistic variables 'above and beyond phonology'. In effect, analysts had to become much more rigorous and explicit in how they treated the data.

In order to study the linguistic variable a two-step methodological process is required; first, identification of two or more variant expressions of a common underlying form; second, an accountable method for deciding all the possible variants and the contexts in which they occur; third, the source of the data must be accountable too, representing authentic data in a diversity of contexts.

A key principle underlying this method (see also Chapter 1) is 'the principle of accountability' (Labov 1982: 30). This principle is fundamental to variation analysis; it dictates that all occurrences of the target variable must be taken into account, not simply one variant or another. In other words,

> analysts should not select from a text those variants of a variable that tend to confirm their argument, and ignore others that do not. (Milroy and Gordon 2003: 137)

In other words, you must include all non-occurrences as well (Labov 1982: 30). Then, the occurrence of variants can be calculated out of the total number of contexts in which it could have occurred, but did not (proportional analysis; see Chapter 9). Similarly, statistical methods can be used to evaluate and compare different contextual effects as well as to detect and measure tendencies over time. Statistical techniques also permit correlations to be made among social and linguistic features. Still, a critical assumption underlies these procedures – the idea that the variants differ relatively little in terms of their function.

When the linguistic variable lies beyond phonology, the variants may not be similar at all. They may have entirely different lexical sources as well as different histories in the language. For example, the alternations between the *will* future and the *going to* future, as in (4), have distinct verbs as their source, Old English *willan* and the motion

verb 'to go'. Alternation between *was* and *were*, as in (5), derives from two different verbs, present tense *beon* 'to exist' and past tense *wesan* 'to dwell'.

(4)
I think she's ***gonna be*** cheeky . . . I think she'***ll be*** cheeky. (YRK/x)

(5)
There ***was*** always kids that ***were*** going missing. (YRK/h)

Such dissimilarities make it impossible to derive the variants from any meaning-preserving grammatical rule. Even the apparently mundane variation between *come* and *came*, as in (6), can be traced back to upheaval in the strong verbs of English in which varying vowel sounds within the verb stem produced different pronunciations of 'come'.

(6)
And Laura ***come*** in at five pound odd . . . I ***came*** in on the Friday . . . (YRK/J)

'Furthermore, the variant whose written form is *come* is much older than *came*. This highlights another issue – the variants may have entirely separate histories in the language not explicable on purely structural terms.

In the case of variables functioning at the level of discourse or pragmatics, the notion of semantic equivalence becomes even more problematic. For example, the variable constructions in (7), which include subject drop (7a), use of *like* (7b) and post-posing in (7c), may be considered semantically distinct.

(7)
a. ***Ø*** used to rent a house with er my mother's sister and cousins. Yeah, so ***we*** used to rent this big house . . . (YRK/w)
b. ***Just like*** little carriages, yes. Yes, ***just Ø*** little tiny things, yes. (YRK/9)
c. ***I was*** terrible, really . . . Very selfish, ***I was***! (YRK/9)

Such cases are problematic for the original grammatical formalism of the variable rules as variants arising from a common underlying form, transformed by some rule of grammar.

In theory, no two forms can have identical meaning, but in practice two different forms can be used interchangeably in some contexts even though they may have distinct referential meanings in other contexts. In fact, you are dealing with at least two different levels of meaning: 1) comprehensive meaning, which takes into consideration every possible inference; and 2) meaning as it is used in the speech community.

While the first is subject to idiosyncratic interpretation and an infinite range of potential meanings, the second is by definition a consensus that is shared and relatively constant. The claim is that meaning in the latter sense should adhere to a narrower interpretation, and be restricted 'to designate the coupling of a given sentence with a given state of affairs' (Weiner and Labov 1983: 30). Indeed, the definition of the linguistic variable may be defined as the task of 'separating out the functionally equivalent from the inferentially possible' (Weiner and Labov 1983: 33). In other words, a foundational task in variation analysis is to 'circumscribe the variable context', the painstaking task which requires the analyst to 'ascertain which structures of forms may be considered variants of each other and in which contexts' (Sankoff 1982: 681).

RE-EXAMINING THE DEFINITION OF THE LINGUISTIC VARIABLE

When analysts first started analysing morphosyntactic variables, they borrowed the notion of semantic equivalence from the model of transformed and untransformed sentences in theories of grammar from the late 1960s (Weiner and Labov 1983). The problem of working out the common underlying grammatical basis for variants embroils the analyst in decisions about underlying and derived forms, which may differ depending on the theory of grammar, which, at the time when this first became an issue, was transformational-generative grammar. Variable rules beyond phonology did not work in this model for two main reasons. First, transformational rules were supposed to be meaning-preserving. However, with morphosyntactic variables this could not easily be defended in any theory of grammar, variationist or other. Second, forms which seemed to be equivalent to each other could often not be derived by the same transformational path.

However, these problems are not intrinsic to the nature of the linguistic variable itself, but are the result of the formalism in which they are embedded. As Sankoff and Thibault (1981) argued, the method of variation analysis obviates these problems. According to standard methodological procedures, the first step is the observation that two (or more) forms are distributed differentially across a community or within the discourse. In other words, the variationist method can only begin when the analyst is convinced that she is dealing with a bona fide variable. Indeed, the particular nature of the underlying form, or even its existence, is irrelevant (Sankoff and Thibault 1981).

You might ask, 'How can this be?' It comes back to the distributional facts of language. The advantage of variation analysis is working with

real data, often from representative samples of communities, and from scrutiny of hundreds and perhaps thousands of instances of the linguistic variable. With this type of data on hand, the distributional facts about language use can be employed for understanding the nature of variation.

In the late 1970s and early 1980s studies of variation above and beyond phonology were breaking new ground. It is not surprising, then, that the operational definition of the linguistic variable was challenged (e.g. Lavandera 1978, 1982). The analytic method needed to be extended, revised and documented.

Sankoff and Thibault's study of weak complementarity demonstrated that the linguistic variable need not be semantically equivalent. Instead, discourse equivalence, or functional equivalence, was found to be the relevant criterion. Indeed, they argue that in many cases 'the most we will be able to say is that the proposed variants can serve one, or more generally, similar discourse functions. We cannot even require that they be identical discourse functions' (Sankoff and Thibault 1981: 208).

So how is one to recognise a linguistic variable, then? Even once you think you have found one, how can you be sure it is a good one? I now turn to exemplifying this pursuit in practical terms.

RECOGNISING THE LINGUISTIC VARIABLE

The linguistic variable can exist at virtually any level of the grammar, ranging from phonetics to discourse, from phonology to syntax, as in (8) (Wolfram 1993: 195):

(8)
a structural category, e.g. the definite article, relativisers, complementisers
a semantic category, e.g. genitive *-s* vs *of* genitive, periphrastic comparative *more* vs synthetic *-er*
a particular morpheme category, e.g. third person singular present tense suffix, the *-ly* suffix on adverbs
a phoneme, a systematic or classical definition of a unit, e.g. [θ] in English
a natural class of units in a particular linguistic environment, e.g. final stop consonant clusters in word-final position, Canadian Raising the process by which the onsets of the diphthongs /ay/ and /aw/ raise to mid-vowels when they precede voiceless obstruents (the sounds /p/, /t/, /k/, /s/ and /f/)
a syntactic relationship of some type, e.g. negative concord, passive vs active
permutation or placement of items, e.g. adverb placement, particle placement
a lexical item, e.g. *chesterfield* vs *couch* vs *settee*

In this way, the linguistic variable is an abstraction. The varying forms must exist in some linguistically meaningful subsystem of the grammar. The linguistic variable must also have another important characteristic. It must co-vary, i.e. correlate, with patterns of social and/or linguistic phenomena.

A linguistic variable is more than simply a synonym, and more complex than simply two ways of saying the same thing. It must also have qualities of system and distribution as well, as in (9), even if these are only revealed by analysis:

(9)
a. synonymy or near synonymy (weak complementarity)
b. structurally embedded, i.e. implicated in structural relations with other elements of the linguistic system, e.g. the phonemic inventory, phonological space, functional heads, grammatical subsystems, etc.
c. correlation with social and/or linguistic phenomena

The fact of the matter is that the onus is on the analyst to provide defensible arguments to demonstrate relevant social and linguistic correlations. In other words, the proof of whether or not a linguistic variable is a linguistic variable is in the pudding.

In sum, early controversy over the extent to which the linguistic variable could be applied to all levels of grammar was really a developmental phase in variation analysis when definitions were being refined and improvements to the methodology were ongoing. Lavandera (1978) correctly pointed out that the linguistic variable, as it had originally been defined, could not be extended to variables above and beyond phonology. However, the research paradigm quickly caught up. Weiner and Labov (1983), Sankoff (1973, 1980), Sankoff and Thibault (1981) and Laberge (1980) demonstrated through detailed methodological argumentation that the linguistic variable need not be confined to cases in which the variants necessarily mean precisely the same thing. Instead, the linguistic variable may have weak complementarity across the speech community, i.e. functional equivalence in discourse. This malleability implicates the role of the linguistic variable in linguistic change (Sankoff 1982: 681–5, 1988b: 153–5, Sankoff and Thibault 1981).

LINGUISTIC VARIABLES AS LANGUAGE CHANGE

How can a linguistic variable involve variants that have no structural relationship or one-to-one equivalence? The answer has to do with

how language changes. Linguistic change does not always occur gradually from one closely related form to another. Instead, language change may proceed by cataclysmic means:

> by forcible juxtaposition of grammatically very different constructions whose only underlying property in common is their usage for similar discursive functions. (Sankoff and Thibault 1981: 207)

Consider a number of examples. *Going to* and *will* are variants of future temporal reference in contemporary English, despite different sources in separate lexical verbs. In earlier times (and perhaps even today) the simple present tense varied systematically with the progressive, e.g. *the kettle boils* vs *the kettle is boiling*, *I love it* vs *I'm loving it*, etc. The relativiser *that*, a complementiser, often varies with *who*, a pronoun.

If one form appears to be replacing the other, either in time or along some socioeconomic or demographic dimension in the community (Sankoff and Thibault 1981: 213), then this may be an indication of change in progress. For example, if a variant is correlated with age, this may be evidence of ongoing evolution of a subsystem of grammar.

The application of variation analysis to formal models of grammatical change was foreshadowed in research in the early 1980s, long before variation analysis was explicitly applied to grammaticalisation theory per se (e.g. Poplack and Tagliamonte 1998, 2001). Sankoff and Thibault (1981) argued that when discourse alternatives coexist over time we may expect this equivalence to eventually become grammaticalised, i.e. functional analogues will become syntactic analogues. They speculated that the criterion of weak complementarity could be used as a diagnostic for stages in the development of forms. The progression of such change might be outlined as follows:

1. An innovation is introduced, it takes on the form of a discourse marker having some attentional or accentuation purpose.
2. The form gradually loses some of its original emphatic qualities.
3. Semantic distinctions gradually become neutralised.
4. Forms grammaticalise and take on the conventional characteristics of a linguistic variable.

Such an approach makes important and testable predictions for grammatical change, as in (10).

(10)
Predictions for grammaticalisation

Early stage	Later stage
Semantic constraints	Neutralisation of semantic constraints

Much more work needs to be done in this area. The challenge is to find the right set of circumstances, a diagnostic variable, and then to test the hypotheses of change. Variation analysis is ripe for research of this kind, and it appears to be a welcoming new frontier for future research:

> a fuller integration of sociolinguistic and developmental research with research on grammaticalization still remains to be worked out. (Hopper and Traugott 1993: 30)

The next question is: How do you choose which variable to study?

SELECTING A LINGUISTIC VARIABLE FOR ANALYSIS

Beyond the motivation to study something that interests you, what are the qualities that you should be looking for when choosing a linguistic variable? Wolfram (1993: 209) notes that 'selecting linguistic variables for study involves considerations on different levels, ranging from descriptive linguistic concerns to practical concerns of reliable coding'. These may seem overwhelming at first, but as you get the hang of it these decisions keep the process vibrant and intriguing.

IDENTIFY POTENTIAL VARIABLES

The first task is to identify potential variables in language. Faced with your data, where do you start? Students often ask me, 'What do I look for?' This is an entirely practical issue. The place to start is to take a long, hard look at your data. As discussed earlier in Chapter 1, language materials, of any type (e.g. written, spoken or otherwise), offer you a wide range of variables for investigation. All you have to do is find them. In the first instance, simply listen, read or look. What is different? What is interesting? Take notes about the things you observe. In some cases they may be structures that are not 'standard' English, or perhaps structures that are different from what you are familiar with in your own variety of English. In fact, when linguistic variables involve dialectal, informal, or non-standard variants they are a lot easier to spot. You tend to notice things that are different from your own idiolect. In other cases, you will need to focus intently on the flow of forms and structures in the discourse because the variables will slip by without you even realising they are there. Many linguistic

variables in contemporary varieties of English, for example, comprise variants which are more or less acceptable in the language, with little associated stigma or affect. Variation is everywhere; you just have to notice it. Sometimes it is right under our noses, as in (11).

(11)

You ***got to*** breathe and have some fun . . . We ***must*** engage and rearrange.
(Lenny Kravitz, 'Are you gonna go my way')

A corpus collected using standard sociolinguistic interviewing typically contains one to two hours of speech per individual, which translates to approximately fifty pages of double-spaced words. Such materials will typically be replete with potential variables. In (12), we have an excerpt from a transcription of Mel, a 40-year-old male in the York English Corpus who works as a computer software trainer. The interview is very relaxed and he presents himself as an easy-going ex-hippie. This excerpt tells the story of how he quit one of his previous jobs. It involves a dramatic exchange between himself and the boss. Bold, underline and italics represent variants of the linguistic variables I will discuss momentarily. Italics represent potential linguistic variables. What I mean by 'potential' is that variants occur that the analyst may infer will vary with other forms in the larger context.

(12)

York English Corpus, Male, age 40
. . . So . . . *sort-of-like* ***jus***' sat in Fibbers, ***havin'*** a ***pint*** and the phone rang, and it was my boss. . . . Oh! Oh, it's- ***tol'*** everybody I'd gone *t'pub*, they knew where to find *me* if they wanted *me*, *you-know*. And er, so the phone rang and it was the boss, *you-know* and she said, 'If- w - - what are you ***doing***?' So I said, 'Well I'm ***havin'*** a beer.' What do you think? 'Er, what about- . . .' Can't think of the name of- *the* guy's name, 'What about *this* guy's manual?' You-see. So I said, 'Well I'll do what I normally do.' You-know, Said, 'I'll do it at '*ome* tonight. It'll *be sorted*.' You-know, I said, '*Have* I ever let you down . . . before?' So she said, 'No.' So I said, 'Well, why are you ***hasslin'*** now?' So she said, 'Well, I want ***something*** on my desk by five-o-clock.' *You-see*, well, 'You'*ve got no* chance.' 'Well when can I see it?' So I said, 'Don't worry, there'll be ***somethin'*** on your desk by nine o'clock tomorrow.' Put the phone down. That night w- - was- a few of us from work . . . ***goin'*** out for a drink, so we're all sat over in the Red-Lion and *like* all these horror stories start ***comin'*** about, about *you-know*, how Joanne's *treat[?]ed* ***differen[?]*** ones of them you-know, and *shit* on them *and what have you*. 'Cos *it was like*, there's two bits. There's a ***recruitmen[?]*** bit and the ***training*** bit. And *I-mean* I was *sort-of-like* ***tucked[t]*** away upstairs by *myself* so I didn't get to see much of what ***wen[?]*** on downstairs. And *they were like* all- we were all sat in the ***pub[u]*** and everybody's ***bitchin'*** about *this* woman, *you-know* and I thought, 'Well I don't want to work with someone like *this*.' You-know, and I ***jus***' said so, I said, 'That's it, I'm ***'anding*** *my* notice in tomorrow.' And *you-know* they'*re* all ***goin'*** *like*, 'Nah,' *you-know*, 'you won't, you won't.' ***Followin***' ***mornin***', um, c-you-know *I-mean*

I'd ***told***[d] 'em about w- - *this* phone call. *You-know* and then when she'd said *like*, everybody had said oh, I though, 'Well '*ang on a minute*, I've said there'd be ***somethin'*** on her desk by nine-o-clock tomorrow ***mornin'***, it will be *my* re - - be *my* notice.' You-know everybody's ***goin'***, 'Oh you won't you won't.' ***Follow in' mornin'*** I got *up*[ʊ], shirt and tie on, suit as normal, ***tootled*****[d]** around the corner, ***walked[t]*** into the office, and I said 'Joanne, you wanted ***somethin***' on your desk by nine-o-clock, there's *my* time sheet, I quit.' . . . And ***walked[t]*** out. And you could ***jus***' see everybody's face like drop. It's like . . . he's done it!

Even in this small excerpt, approximately three minutes of a two-hour interview, there are many features that hold promise for investigation. A number of linguistic variables can be authenticated. What I mean by this is that the alternatives are both visible.

Variable (ing) and variable (t,d)

Two variables readily apparent in this excerpt are variable (ing) and variable (t,d). Note that this excerpt has been embellished from the transcription file, with an indication of the actual pronunciation of the forms for illustration purposes. In fact, these are two of the most widely studied variables in the history of variation analysis. Take a closer look at each of the instances of these variables. The words in which they occur have been bolded, italicised and underlined for easy visibility. I have also indicated which of the phonological variants was produced in each case. The words containing variable (ing) and (t,d) are listed in (13) and (14) respectively.

(13)
Variable (ing)
havin', doing, *havin'*, *hasslin'*, something, *somethin'*, *goin'*, *comin'*, training, *bitchin'*, 'anding, *goin'*, *followin'*, *mornin'*, *somethin'*, *mornin'*, *goin'*, *followin'*, *mornin'*, *somethin'*

(14)
Variable (t,d)
jus', pint, *tol'*, different, recruitment, tucked, went, *jus'*, told, tootled, walked, walked, jus'

How many of each variant occur in each variable set? For (ing), notice that the standard variant [ŋ] occurs only four times. For variable (t,d), there are four examples of the non-standard, zero form. The semi-weak verb *told* (in line 2), and monomorpheme *just* (lines 1, 21, and 31) exhibit simplification of the consonant cluster. In other words, this speaker uses mostly non-standard [n], but standard [t,d] forms in his speech. In the full studies of both these variables, these idiolectal tendencies hold across the broader sample of York English

(Tagliamonte 2004, Tagliamonte and Temple 2005). Overall there is relatively frequent use of the standard variant of variable (t,d), i.e. realised clusters, compared to other varieties. In contrast, the standard variant of variable (ing), i.e. the velar variant, is quite rare.

A multitude of other interesting and potentially variable forms are evident – some phonological, (15), and others morphological and syntactic, (16). These have been italicised in the excerpt.

(15)

Phonological		
a. definite article reduction	gone *t'*pub	line 2
b. variable (h), dropping	*'ome*	line 7
	'anding	line 21
	'ang	line 25
c. variable (t)	*trea*[ʔ]*ed*	line 15
d. variable (U)	*pub* [pʊb]	line 19

(16)

Morphological and syntactic		
a. *of* vs *'s* genitive	the name *of -the* guy*'s* name	line 6
b. agreement	there*'s* two bits	line 16
c. subject drop	Ø put the phone down	line 12
d. zero definite article	*following mornin'*	line 23, 27
e. possessive *have got* vs *have*	you*'ve got no* chance	line 10

Many discourse/pragmatic features are evident as well, as in (17):

(17)

Discourse/pragmatic		
a. extension particles	*and what have you*	line 16
b. quotatives	*said*	line 4, 6
	thought	line 20
	going . . .	line 22, 27
	it's like	line 31
c. discourse *like*	*it was like . . .*	line 16
	like drop	line 31
d. discourse markers	*you know*	line 3, 4, 7, 14, 15, 20, 21, 22, 23, 24, 26
	I mean	line 17, 23
	you see	line 6, 10
e. discourse *so*	*so the phone rang*	line 3
	so I said	line 6
	so we're all sat	line 13

Of course, in such a small excerpt of material most of these potential variables cannot be authenticated. In other words, only one variant is actually present. You cannot be sure that the linguistic feature

in question is variable in the data from the available evidence. However, if you know these variants participate in alternation with other forms, then the presence of even one of the variants is a good indication that the other may be present as well. Further examination of a greater portion of the data for this speaker would confirm which are variable and which are not. Nevertheless, the sheer number of possible features for study is quite remarkable.

Other features of note are morphosyntactic and lexical features that stand out nationally, regionally and locally, as in (18).

(18)

a. *we're sat . . .*	vs	we're sitting
b. *it'll be sorted . . .*	vs	it'll be fixed/worked out, etc.
c. *tootled around . . .*	vs	walked
d. *hasslin' . . .*	vs	bothering/bugging, etc.

Faced with such a data set, the analyst must decide which variable to tackle for a fully fledged analysis. Which one would you choose?

Notice in (12) that variable (ing) is quite frequent, occurring nearly once per line, for a total of 20 times. Variable (t,d) occurs 11 times. It is not surprising that these two variables have been so often studied in the literature. They are easy to spot and easy to find. Both characteristics are ideal criteria for selecting a linguistic variable.

In fact, some linguistic variables are better candidates for variation analysis than others. Variable items which lack systemic, linguistic foundations such as variable realisations of words like 'yes', (19a), 'because', (19b), or performance anomalies, (19c–d), may not be ideal for variation analysis.

(19)

a. ***Yes*** it has, very tiny. . . . ***Yeah*** they're not- they're not that big. (YRK/™)
b. ***'Cos*** the atmosphere up there's different as well ***because*** um everyone's doing exams. (YRK/U)
c. We ***just go***- really ***we'd um***- we'd just go out . . . (YRK/™)
d. The ***b--*** the ***boys*** from Brigg were um- ten of their team were- (YRK/U)

A number of criteria can guide the analyst in choosing a 'good' linguistic variable for analysis. Ideally, you want to select a variable that is interesting and relevant, both to you and within the field. But, in practice, this goal must inevitably be balanced on practical grounds.

Frequency

Linguistic features that are rare, either because of the relative infrequency of the structure or because of conscious suppression in an

interview, may not be good candidates for analysis. They may be interesting linguistically, dialectally fascinating and critical for a comprehensive descriptive profile, but if they do not occur with sufficient numbers they can hardly be tabulated in a study of variation. Phonological variables are usually more frequent, while grammatical structures are rarer. Discourse features may be remarkably frequent or virtually absent depending on the variety under investigation, age of the speaker, etc.

Sometimes features occur extremely frequently, but cannot be ideal variables because the context of variation is questionable. This arises most obviously in the case of discourse-pragmatic features, where only one variant is overt in the discourse. But what is its alternative? Where *can* it occur, but did not? In contemporary English, features of this type are plentiful, including *like*, *anyway*, *so*, etc. My students always want to study these features. What they do not realise is the study of these forms using variation analysis is a very complex and difficult enterprise. Defining the variable context requires painstaking treatment of the data and advanced knowledge of syntax because the feature must be defined structurally in order to assess its function in the phrase structure (see D'Arcy 2005).

It is possible to structure interview schedule/questions to elicit specific types of constructions. For example, talking about past time will enhance the occurrence of past tense forms; talking about habitual activities will enhance the occurrence of habitual tense/aspect forms; and getting informants to tell you stories will enhance your ability to get quotatives. However, you may not know in advance which feature(s) you want to study, or which features may become important to you later on. In sum, not all goals can be achieved in every interview situation. The frequency of different types of variables depends greatly on the type of discourse situation and innumerable other, often uncontrollable, factors.

Tip

One of my strategies for finding a good linguistic variable is to compile an index of my interviews and look closely at the words in the data that occur most frequently (see Chapter 4). Another strategy is to read prescriptive grammars and find cases where alternate forms are mentioned. Another is to simply observe what linguistic variables researchers are talking about and check to see what is happening with those variables in your own data. If it is frequent enough, and the variation is robust enough, it is a good candidate for further investigation.

Robustness

Frequency is not necessarily the choice criterion for selecting a linguistic variable. A further requirement is that there is adequate variation between forms. Linguistic variables which are frequent but have minimal variation are less viable for investigation by this method. Although the structures themselves may be interesting, if the data at your disposal is near categorical (either 100 per cent or 0 per cent), then there is little room for quantitative investigation. If variability hovers at very low or very high levels, differences between variants in independent contexts may be too small to achieve statistical significance. In this case, you may rely on the constraint ranking of factors for comparative purposes (Poplack and Tagliamonte 2001); however, near categorical variables may not have sufficient numbers for even constraint ranking to be informative. In such cases, one of the possible variants may have such marginal status in the data that the variable itself will be unrevealing. If it is a change in progress, it may also be possible that the variable has either 'gone to completion' or is perhaps still so incipient, or so marginal in the data, that it cannot be reliably modelled using statistical methods.

Sometimes very low-frequency items, by their very characteristic of limited status in a variety, can be extremely important. Indeed, Trudgill (1999) argues that 'embryonic' variants may sometimes blossom into rampant change. Something of this nature has occurred in the contemporary English quotative system where a new form, *be like* as in (20), represented only 13 per cent of all quotative verbs in Canadian English in 1995 (Tagliamonte and Hudson 1999).

(20)
I'm like, 'You're kidding? Wow, that's really cool.'
She ***says***, 'What do you think of him?'
I said, 'Well, yeah, he's cute.' (OTT/c)

Yet in the early 2000s it has risen to become the dominant quotative, 65 per cent – as in (21) (Tagliamonte 2005) – a four-and-a-half-fold increase in less than eight years.

(21)
She'***s like***, 'Have you taken accounting?'
I'm like, 'No.'
She'***s like***, 'Have you taken business?' (TOR/I/@)

A low-frequency variable which was well worth investigating was pre-verbal *do* in Somerset English, as in (22) (Jones and Tagliamonte 2004).

(22)
We ***did have*** an outside toilet, just a brick type of thing, you know.
We ***did have*** a flush toilet there. (TIV/e)

Minimal presence of periphrastic *do* amongst the oldest generation and virtual absence amongst the youngest generation meant that this feature was finally dying out of the variety. This study likely represents the last opportunity to discover the grammar of this feature before it disappears for good. Therefore, despite the highly infrequent status of the feature, we decided to study it anyway.

Unfortunately, some obsolescent features in contemporary English are so far gone that they cannot be studied quantitatively at all. This was the case for the *for to* complementiser in British dialects, as in (23). While we attempted to tabulate its frequency and distribution in our data, in the end it was too rare for substantive patterns of use to be revealed in the data (e.g. Tagliamonte et al. to appear).

(23)
a. So the roads were crowded when it was time ***for to*** start. (MPT/v)
b. He'd light a furnace ***for to*** wash the clothes. (TIV/a)

In sum, there may be extenuating circumstances for selecting a linguistic variable where one of the variants has very low frequency. Under most circumstances, however, variation analysis is best suited for a linguistic variable where at least some of the variants occur robustly. This permits a richer, more complex and informative analysis.

Implications for (socio)linguistic issues

Your choice of a linguistic variable should also be dictated by the extent to which it has the capacity to answer timely and relevant questions. For example, linguistic variables that are undergoing change are excellent targets for analysis since they give insights into the process of change itself. Those that implicate grammatical structures reveal details of the syntactic component of grammar. Those that differentiate dialects highlight parametric differences and so on.

Once you have decided which variable you will study, what next? It is time to extract all instances of the variable from your data according to the principle of accountability.

CIRCUMSCRIPTION OF THE VARIABLE CONTEXT

Deciding on precisely how and where in the grammatical system a particular linguistic variable occurs is referred to as 'circumscribing the variable context' (e.g. Poplack and Tagliamonte 1989: 60). This refers to the multitude of little decisions that need to be made in order to fine-tune precisely where alternates of a linguistic variable are possible.

The procedure for inclusion and exclusion of items must be set forth explicitly so that your analysis is replicable. If you do not provide this information, you violate the researcher's obligation to provide enough information for your study to be repeated with reasonable accuracy and hence comparability.

First, you must identify the contexts in which the variants occur. Do each of the variants occur with all speakers? Do certain subgroups use more than others? These questions lead the analyst in identifying the envelope of variation (Labov 1972c). The tricky part is that you must count the number of *actual* occurrences of a particular structure *as well as* all those cases where the form might have occurred but did not. In other words, you have to know 'what is varying with what' (Weiner and Labov 1983: 33). In fact, you must know what the alternative variants are, even when one of the variants is nothing at all. But if one of the variants is zero, as is often the case, how do you spot them?

This is where the task of circumscribing the variable context can present special difficulties. Moreover, depending on the linguistic variable, there will be confounding factors that necessitate the exclusion of some instances, or tokens, of the variable.

CATEGORICAL, NEAR CATEGORICAL AND VARIABLE CONTEXTS

There may be a particular context in which one or the other variant never occurs. This is called a 'categorical context', which means that the variable is realised either 0 per cent or 100 per cent of the time. Such a case must necessarily be excluded from variable rule analysis for the simple reason that it is invariant. This is not to say that categorical contexts are not important. They are. In fact, the contrast between categorical variable contexts are diagnostic of structural differences in the grammar.

However, if the categorical environments were included in the variable rule analysis:

1) the frequency of application of the rule would appear much lower than it actually is,
2) a number of important constraints on the variable contexts would be obscured, since they would appear to apply to only a small portion of the cases, and
3) the important distinction between variable and categorical behaviour would be lost (Labov 1969a, 1972c: 82).

For example, consider variation in the presence of periphrastic *do* in negative declarative sentences in a northern Scots variety, as in (24) (Smith 2001).

(24)
a. I ***dinna*** mine fa taen it. (BCK/a)
b. I ***na*** mine fa come in. (BCK/a)

Smith demonstrated that there were two types of contexts: 1) those that never (or rarely) had *do* absence, third person; and 2) those that were variable, first and second person. While the (near) categorical contexts could be explained on syntactic grounds, the variable contexts were conditioned by lexical, frequency and processing constraints. The divide between these two types of contexts showed the importance of the categorical/variable distinction in the grammar.

How do you circumscribe the *variable* contexts? If the context is 95 per cent or over, 5 per cent or under, these are also transparent candidates for exclusion from the variation analysis (Guy 1988). However, in most analyses there will be a wide range of frequencies across factors. The analyst must be aware of where the variation exhibits extremes at one end of the scale or the other, as these contexts will be critical for explaining the variation.

In other words, the questions to ask yourself as you define the envelope of linguistic variation are these: Does this token behave exceptionally? Does it behave like other tokens of the variable? The major part of circumscribing the variable context is to 'specify where the variable occurs and where it does not' (Weiner and Labov 1983: 36). In so doing, you must provide an explicit account of which contexts are *not* part of the variable context.

The decisions that go into circumscribing the variable context affect the results in very important ways. Be sure to make principled decisions at each step in the process. Even the most sophisticated quantitative manipulations will not be able to save the analysis if you do not do this first (Labov 1969a: 728). In the next section I turn to some practical examples.

Tip

Don't be afraid to falsify your own procedures! Circumscribing the linguistic variable is a process that unfolds as you go and is continually revised nearly right up to the end of the extraction process. I don't know how many times I've had to go back and include a token type because I found later that it was variable. I've also had to go back and exclude tokens that were later found to be invariable. This is all part of the discovery process. But remember to document everything!

EXCEPTIONAL DISTRIBUTIONS

One of the first things to attend to when circumscribing the variable context is whether or not there are contexts in the data that are exceptional in some way. Exceptional behaviour often becomes obvious only as research evolves. Certain exceptional behaviours are part of the knowledge base existing in the literature. It is the responsibility of the analyst to know what idiosyncratic behaviour has been noted in earlier research and to pay particularly good attention to how the variants of a variable are distributed in the data set under investigation. Are the co-varying nouns, verbs, adjectives, etc. behaving comparably? Are different structures, sentence types and discourse contexts the same, or different? Exceptional distributions may occur for any number of reasons and these will differ depending on the variable and depending on what is going on in a data set. This is undoubtedly part of what Labov meant by 'exploratory manoeuvres' (Labov 1969a: 728).

Asymmetrical contexts

It is critical that each linguistic variable be scrutinised for asymmetrical distribution patterns. For example, in a study of verbal *-s* in Early African American English (Poplack and Tagliamonte 1989), we knew that one of its salient characteristics was its use with non-finite constructions (Labov et al. 1968: 165). For this reason, we were looking for cases of verbal *-s* in these constructions in our data. When we did not find any, it was readily apparent we were dealing with a different situation. Similarly, we knew from earlier research that verbal *-s* tended to appear on certain verbs only. Once again, this was a red flag to us to pay attention to the distribution of variants by lexical verb.

Another good illustration of exceptional behaviour that must be taken into account comes from the study of relative markers in English, as illustrated in (25) from a single speaker in the York English Corpus. At the outset, it is extremely important to isolate the restrictive relative clauses. Why? Because in contemporary varieties of English, non-restrictive relative clauses differ on a number of counts from restrictive relatives, and thus cannot be treated in the same analysis. First, non-restrictive relative clauses occur primarily with *which* and *who*, but hardly ever with *that* and zero; second, their semantic function differs; third, non-restrictives are marked off prosodically (as indicated by commas in (25)). Given these characteristics, if non-restrictive relatives were included in a sample of data which included restrictive relative markers, as in the embedded clause in (25), the effect would be to raise the percentage of *which/who* forms and lower the percentage of the others (*that* and zero). Further, the results would not be comparable with other data where only restrictive relative clauses were studied.

(25)

Albert, ***who*** was one of the guys ***that*** I knew from the Bayhorse, got him to do his physics homework for him. (YRK/Σ)

In other words, because non-restrictive relative clauses are nearly categorically marked with 'wh' forms, they are exceptional when it comes to the presence of relative markers, and should not be included in the same analysis as restrictive relatives (see also Ball 1996).

Somewhat the same modus operandi led to numerous exclusions in my study of dual form adverbs (see Tagliamonte and Ito 2002: 246–8). The variation was restricted to adverbs that could take either *-ly* or *-Ø*, without a difference in function. Numerous adverbs had to be excluded which did not permit *-ly*, e.g. *high*, or whose adjectival form (i.e. the zero form) was not semantically related to the *-ly* counterparts, e.g. *shortly*. For example, *directly* in (26a) was excluded because it means 'immediately' in this context. However, the token in (26b) was included because *direct* in this context can alternate with *directly*, meaning 'in a direct way without deviation'.

(26)

a. He drove home ***directly*** after arriving (= 'immediately').
b. 'Cos in those days as well you used to get er milk ***direct*** from a – a- dairy on a morning. (YRK/?)

Sometimes you will not know a priori which contexts are variable and which are not. This is particularly true when you have targeted a

variable which is undergoing change. Your own intuitions may not match what is happening in the speech community. For example, also in my study of dual form adverbs, I adopted a strategy of examining the data itself for evidence of a particular item's variability. This is because the literature and my own intuitions often failed to make the appropriate judgements about potential variability for the adverb (Tagliamonte and Ito 2002: 247). Indeed, a reviewer of the study criticised us for including certain types, as in (27), which he or she claimed were not variable. In the rewrite we had to demonstrate that they were, in fact, variable and, further, that they were non-negligible in number and diffused across a reasonable proportion of our speakers. We used these distributional facts to justify their inclusion in the analysis.

(27)
a. I was an angel, ***absolute***. (YRK/?)
b. I had years of utter misery, ***absolutely***. (YRK/?)

A variable must be investigated in tremendous detail in order to determine which contexts permit variation and which do not. Those that do not must be listed, and reasons for their exclusion explained.

Formulaic utterances

Typical constructions which exhibit exceptional behaviour for linguistic variables are those that have been learned by rote such as songs, psalms or sayings, as in (28a). In addition, metalinguistic commentary, as in (28b), is a context for exclusion since these constructions may be imitative. Therefore, neither (28a) nor (28b) were included in our study of plural *-s* (Poplack and Tagliamonte 1994).

(28)
a. I look up to the *hill**s*** where cometh my help. (SAM/J)
b. And then they say, you know, '*potatoe**s***'. They say '*potatoe**s***'. (NPR/008)

Exceptional distributions also occur in expressions where the individual lexical items have become part of a larger 'chunk'. In the study of verbal *-s*, contexts such as *I mean*, *you know*, *I see* were excluded, as they were invariant (Godfrey and Tagliamonte 1999: 99–100). This is, of course, because they are functioning as discourse markers, not verbs, as in (29a–b). Similarly, in a study of past tense *be* (variable *was/were*), contexts such as in (29c) were excluded (Tagliamonte and Smith 2000: 160).

(29)
a. We'd seen the roses, ***you see***. (YRK/d)
b. Should have made it a bigger thing, ***I think*** (YRK/d)
c. So, I had friends, ***as it were***, from my own environment. (YRK/8)

When the variable under investigation occurs in a context which is anomalous with respect to the variation of forms within it, these are typically removed from the analysis.

Neutralisation

Neutralisation contexts are tokens in which independent processes exist which make the reliable identification of the variant under investigation difficult (or near impossible). In other words, unambiguous identification of the variant is compromised. The simplest case of neutralisation comes from variables which are phonologically conditioned. For example, the juxtaposition of a noun or verb ending in [s,z] and a following word beginning with [s,z], as in (30), precludes being able to identify the segment accurately as the final suffix on the noun/verb or the initial segment of the following word (Wolfram 1993, Poplack and Tagliamonte 1994).

(30)
a. Pop wa**[s] [s]**at there rubbing her arm. (YRK/c)
b. You get**[s] [s]**ick of them if you had too many. (DVN/1/253)

Similarly, in studies of (t,d) deletion, juxtaposition of a word ending in [t,d] and a following word beginning with [t,d], as in (31), makes it impossible to determine whether the final (t,d) or the initial (t,d) of the following word has been removed.

(31)
We were suppose**[d] [t]**o land on the shore. (YRK/K)

Ambiguity

When a linguistic variable involves a grammatical feature whose varying forms implicate different semantic interpretations, the issue of circumscribing the variable context becomes more difficult. Word-final suffixes such as verbal *-s* or past tense *-ed* involve independent processes of consonant cluster simplification which render the surface forms of regular (weak) present and past tense verbs indistinguishable, as in (32):

(32)
She ***liveØ*** right up yonder. (SAM/E)

Verbs in past temporal reference contexts with no marker are ambiguous. They could be instances of uninflected present tense forms or past tense forms with phonologically deleted [t,d]. Including them will obviously skew the proportions of *-s* presence one way or another. Only forms for which past reference can be firmly established should be included. Past tense readings can often be inferred, for example, from adverbial or other temporal disambiguating constructions, as in (33a), as well as other indicators, as in (33b).

(33)
a. He ***liveØ*** with mama thirty, thirty-two years . . . (ESR/Î)
b. There was a pal ***liveØ*** there. (YRK/®)

Other processes may also render the function of a variant indistinguishable from another. For example, in (34) it is impossible to determine whether the sibilant consonant represents the plural suffix followed by a deleted copula, or a zero plural followed by a contracted copula.

(34)
Them *thing**[z]*** a bad thing. (NPR/4)

Some contexts may be inherently ambiguous. For example, in a study of past tense expression, verbs with identical present and past tense forms such as 'put, set, beat' would not be included because there is no variation one way or the other, as in (35).

(35)
a. past tense
That was before Tang-Hall was built you-see, they ***put*** in sewerage drain from Heworth, the top water and then they ***put*** in- then they got started building. (YRK/¥)
b. present tense
. . . things what you ***put*** your tea in. (YRK/¥)

Another source of ambiguity is when nothing in the context permits an unambiguous interpretation of the form's function. For example, in (36) you cannot tell whether the noun is plural or singular. Therefore, neither of these tokens should be included in an analysis of plural nouns.

(36)
a. Just behind the ***tree***. (SAM/B)
b. I ain't gonna tell no ***lie***. (ESR/Y)

In sum, many contexts may seem to be part of the variable context but are not. Sometimes you may not know they present a problem until

much later. This does not matter. It is more important to include things than not include them, because it is way easier to include more tokens while you are extracting the data than to have to go back and get the ones you missed later on. In fact, excluding certain types of tokens from the data file is simple, as long as they have been treated uniquely in the coding system. I will tell you more about this in Chapters 8 and 10.

Ensuring functional equivalence

With morphosyntactic variables, following the criterion of 'functional equivalence' is often not straightforward. You must be particularly mindful that each variant is an instance of the same function.

The study of tense–aspect features in variation analysis has been particularly helpful in outlining procedures for excluding contexts which do not meet the criterion of functional equivalence. Tense–aspect features are often involved in longitudinal layering of forms in the grammar, in which only a particular subset may be implicated in variation of the linguistic variable under investigation. For example, the study of future temporal reference involves variation in the forms *will* and *going to*. However, different forms of *will* (e.g. *won't*, *'d* and *'ll*) may also denote other (non-future) temporal, modal and/or aspectual meanings. Therefore, any study of future time must restrict the variable context to include cases of *will* that make predictions about states or events transpiring after speech time. This involves identifying and excluding all forms that involve other semantic readings: 1) forms having a modal rather than temporal interpretation, as in (37a); 2) counterfactual conditions that are hypothetical not temporal, as in (37b); or 3) forms denoting habitual action in the present or past, as in (37c).

(37)
a. And today, I ***wouldn't do*** that for the queen . . . (GYE/<)
b. If it ***was*** up to me, I'd have fish on Sunday. (NPR/a)
c. And we ***would go*** hitting each other brothers and then we ***would fight***. (NPR/f)

By strictly circumscribing the contexts to those that are temporal and that make reference to future time, the variants included in the analysis are pertinent to the study of grammatical change in the future temporal reference system.

Repetitions

Tokens which occur directly after another in sequence as false starts or performance errors are typically not included in a variation analysis.

For example, in (38), only the first of the repeated tokens was included in the data file for these variables. Inclusion of repeated tokens would add a disproportionate number of instances of the same form.

(38)
a. And then ***funny*** enough, ***funny*** enough, I think in one year four of us got married. (YRK/?)
b. So they'd ***played*** one short- they'd ***played*** one short. (YRK/π)

Natural speech anomalies

As with all naturally occurring speech, accurate interpretation of any part of the discourse may on occasion be impossible. Intrinsic characteristics of oral discourse like false starts, hesitations, ellipsis and reformulations, as in (39), often lead to difficulty in interpretation. Any unclear or ambiguous contexts should be excluded from the analysis.

(39)
a. And there's another new one in this week ***who***- (CMK/t)
b. And um, it ***was*** *very*– (YRK/c)

IMPOSING AN ANALYSIS

In circumscribing any variable context, you must be aware that your decision-making process may impose an analysis on the data from the outset. A good example of this comes from the study of variable (t,d) in African American Vernacular English (e.g. Labov et al. 1968, Wolfram 1969, Fasold 1972) and then, later, in Guyanese Creole (Bickerton 1975). Part of the variable context involves suffixal (t,d) alternating with bare verbs (i.e. no suffix) in contexts of past temporal reference, as in (40a). Another part involves past marking of strong verbs, alternating with their base forms, also in contexts of past temporal reference, as in (40b).

(40)
a. That's got how many years since they ***killØ*** Papita? Yes, since they ***kilt*** him. (SAM/F)
b. I don't know where they ***came*** from, but anyhow they ***came*** there, they ***begin*** to work. (SAM/J)

Bickerton criticised early studies by suggesting that, if those studies had considered creole categories, such as distinctions of aspect, it would be revealed that the zero-marked verbs resulted, not from deletion of English morphemes, but from a pattern of overt and zero marking peculiar to creoles. In these grammatical systems the zero form actually encodes a different function, a particular aspectual reading.

One way to handle this type of pitfall is to configure your data to allow for different possibilities of analysis. For example, in Tagliamonte and Poplack (1993) we set up the coding system to test for both a creole and an English underlying grammar. No one analysis can claim to be the most accurate; however, a defensible and replicable analysis provides a sound foundation for future research.

THE TYPE–TOKEN QUESTION

The type–token question is whether to include frequently occurring items every single time they occur, or include only some (Wolfram 1969: 58). Such a strategy is particularly relevant for phonological variation where the inclusion of frequently occurring words with exceptional distribution patterns may distort the results. The best example I can think of is a recent study of dialect acquisition in young children (Tagliamonte and Molfenter 2005). The focus of investigation is variable (t) with variation amongst [t], [d] and [ʔ]. In the data, the children, aged 2–5, used the lexical item *little* extremely frequently, as in (41).

(41)
Mum, but we need- ***little*** holes. Why do we need ***little*** holes in it? Can I put ***little*** holes in it? Shaman can I put ***little*** ***little*** holes in? (KID/1)

A standard approach to such a situation is to restrict the number of tokens per speaker, e.g. five tokens per hour of recording per child. However, in the study of acquisition, frequency of forms is critical. In order to model this effect on acquisition it would be necessary to include *all* the forms. In this study we opted for an all-or-nothing strategy by devising a coding schema (see Chapter 6) that enables us to include only five tokens per hour per child or all of them. Time will tell which method supplies a better explanation for the data.

The type-token question may have varying implications depending on the level of grammar under investigation and/or the particular variable targeted. While restricting the number of lexical items in a phonological analysis of variation may be defensible, the same decision might be less so in a study of syntax. The analyst must make a choice as to how her own study will proceed. Whatever the decision, it should be transparent enough for comparison with earlier research as well as future replications. Procedures for how the type-token question is resolved differ across studies and, unfortunately, in many the decisions have not been made explicit in published works. To date, the

relevance of type-token decisions has not, to my knowledge, been fully explored in the published literature.

ILLUSTRATING LINGUISTIC VARIABLES

A requisite component of a variation analysis is to illustrate the linguistic variable. At the beginning, it is important to substantiate the crucial characteristics of equivalence and distribution as well as intra-speaker and inter-speaker variation. In the ideal situation you will find a 'super token': alternation of variants by the same speaker in the same stretch of discourse. Examples of variable verbal *-s* from Samaná English (Poplack and Tagliamonte 1989a: 49) show that both *-s* and zero occur in the same speaker. Examples (42a–b) are uttered by speaker 'E'.

Third person singular

(42)
a. And sometimes she ***go*** in the evening and ***come*** up in the morning. (SAM/E)
b. She ***goes*** to town every morning and ***comes*** up in the evening. (SAM/E)

Tip

Whenever I construct a handout I always look for the most interesting, funny, informative examples I can find in my data. The reasons are: 1) to convey a sense of what the variety under investigation is like; and 2) if the audience is bored, they can at least enjoy the data!

Examples of variable adverbial *-ly* from York English, as in (43) (Tagliamonte and Ito 2002), show that both *-ly* and zero occur in the same speaker as well as in the same stretch of discourse.

(43)
I mean, you go to Leeds and Castleford, they take it so much more ***seriously*** . . . They really are, they take it so ***serious***. (YRK/T)

Providing examples of intra-speaker variation is important because it demonstrates that the linguistic variable under investigation is endemic to individual sample members, not simply the result of amalgamating data from speakers who are categorical one way or another.

Cross-variety comparisons illustrate that variation exists within individuals *and* across the communities under investigation. In (44a) you see intra-speaker variation for African Nova Scotian English in

rural Nova Scotia, Canada, and in (44b), for Buckie English in rural Scotland (Tagliamonte and Smith 2000).

(44)
a. And we ***was*** the only colour family. We ***were*** just surrounded. (GYE/l)
b. We ***were*** all thegither . . . I think we ***was*** all thegither. (BCK/h)

Similarly, example (45) illustrates variable verbal *-s* in third person plural in Samaná English and Devon English (Godfrey and Tagliamonte 1999).

(45)
a. They ***speak*** the same English. But you see, the English people ***talks*** with grammar. (SAM/G)
b. Yeah they ***drives*** 'em . . . They ***help*** out. (DVN/d)

SUMMARY

Where does all this leave you with regard to defining the linguistic variable? The main thing is simply 'know them by their colours'. In other words, the onus is on the analyst to determine and defend the linguistic variable under investigation. If the variable is bona fide, this should become evident during the investigation. This means establishing at the outset that the linguistic variable is authentic, meeting the criteria of 1) functional equivalence; 2) distribution and 3) structural embedding. These criteria are often outlined in research papers as part of the methodology section. As part of the process of *doing* variation analysis, data anomalies may arise, further observations may become apparent and correlations may reveal themselves. Such discoveries can then be incorporated into the analysis, sometimes becoming part of the story. Indeed, as the field has evolved, circumscribing the variable context has become an important starting point and, as Labov says (to appear), it is an important end point too.

In sum, the systematic study of competing forms of variation analysis requires not only the identification of these forms, but also the individual contexts in which differences between them are neutralised. This, in turn, leads to the interpretative component of variation analysis, i.e. deciding how to circumscribe the context and identifying the places in which variation between forms for the same function may occur. I turn to this phase of research in Chapter 6.

Exercise 5: Locating and circumscribing a linguistic variable

One of the key measures of success in the study of language variation and change is to locate an appropriate linguistic variable to analyse. In this exercise you will pay particularly close attention to the data you have targeted and, based on your own observations of variation in your data (as you experienced with Exercise 2), choose the linguistic variable that you would like to study.

The variable should be relatively frequent in the data and have linguistic and/or sociolinguistic implications.

You must establish that the linguistic feature you choose is a *bona fide* linguistic variable, i.e. a linguistic feature which can be *shown* to co-vary systematically with some features of the linguistic or extralinguistic environment.

Your report should include the following:

Identification of your variable

What is it? How many variants are there? What are they? Which are standard? Which are non-standard/dialectal? Describe them and provide examples. If you can find a 'super-token', that is ideal.

Definition of the variable context

Include a precise definition of all contexts which will be included in your analysis.

Exclusions and exceptional distributions

Exclude any forms which are not part of the variable context:

* invariant forms (e.g. a context that is always one variant or the other)
* exceptional distributions (e.g. metalinguistic commentary, quoted speech, etc.)
* ambiguous contexts (e.g. false starts, neutralisation, etc.)
* forms that do not have the relevant function

Illustrate each of these and justify why they should be excluded.
Read sections entitled 'Circumscribing the variable context' in the following:

Godfrey, E. and Tagliamonte, S. (1999). **98–100**.
Poplack, S. and Tagliamonte, S. (1989). **47–84**.
Sankoff, G. and Thibault, P. (1980). **315–30**.
Tagliamonte, S. (1998). **159–61**.
Tagliamonte, S. and Hudson, R. (1999). **154–7**.