8 Variation and change

Gregory R. Guy

8.1 Introduction

Like most human activity, language does not fit neatly into the analytic boxes that observers often use to segment, categorise, and theorise about the subject. Whether those boxes are called features, phonemes, or syntactic structures, or rules, constraints, or principles, the facts of language always slop over the edges or ooze from one into another. The customary approach in linguistics is to treat this mismatch between categories and facts as 'linguistic variation' – but we should be clear that doing so effectively privileges the analytical categories over the empirical substance. Variation, as traditionally understood, involves single categories being mapped onto variable realisations, as if the categories were primary and given – platonic ideals existing on a higher, purer, plane, that are only imperfectly reflected in the muddy reality of speech. An alternative view, in which natural language in all its richly variegated glory is primary, and the analytical categories are as yet imperfect theoretical constructs that provide only a crude model of reality, is rarely considered. As a healthy terminological corrective, perhaps linguists should consider thinking about variation as highlighting the problem of 'theoretical inadequacy'.

Nowhere is this lousy fit between theoretical models and variable facts more evident than in the treatment of language change. Since Saussure, linguistic theory has for the most part assumed the irrelevance of diachrony in the construction of formal theory, producing as a consequence static models that not only fail to accommodate change, but actually appear to exclude it as a logical possibility. Theoretical models are designed to be self-contained systems, supported by their internal structure and logic, covering a strictly defined terrain (from which diversity is excluded). Such theories are like buildings and, as we know from experience, buildings do not evolve organically; rather, they change by getting completely or partially demolished and replaced. Consequently, such theories make change seem anomalous, or impossible, and in any case, located outside of theory. And yet, linguistic reality obstinately refuses to accommodate to these models, and all languages go on changing continuously all the time. What's a linguist to do?

The resolution of these contradictions lies in abandoning the theoretical assumptions that inhibit a proper treatment of linguistic variation and change. Since the Neogrammarians, the main stream of theoretical development in linguistics has been enchanted with the idea that 'exceptionlessness' (Neogrammarian *Ausnahmlosigheit*) is an essential trait of valid linguistic generalisations; since Saussure, variability has been defined as lying outside the linguistic system, external to *langue*, competence, and grammar. But an alternative model exists which avoids these anti-empirical assumptions, in which valid generalisations may be non-categorical, and variation may be seen as systematic and internal to grammar. A path-breaking formulation of this position is found in Weinreich, Labov and Herzog 1968. These scholars enunciate two principles that are foundational to this alternative: orderly heterogeneity is the principle that variation is not equal to chaos, but may still contain system and order, and inherent variability is the principle that variability is intrinsic to language, effectively perceived, processed, and produced by all speakers, and therefore lies within competence, and hence within grammar.

With this change of assumptions, it becomes clear that variation and change are essentially one and the same phenomenon. The speech community, and the mental grammars of speakers, encompass and manipulate linguistic differences at all times. No two speakers have identical grammars and linguistic repertoires, and no single speaker has a completely homogeneous and invariant grammar. Therefore, to say or understand anything at all, a language user must be able to deal with difference, with 'variation'. Unsurprisingly, the particular patterns of difference fluctuate across time, just as they fluctuate across speakers and social situations. Therefore, variation is the synchronic face of change, and change is nothing more than diachronic variation. Indeed, the historical record, along with studies of change in progress, make it clear that there is no such thing as change without variation: all changes pass through periods of time during which outgoing and incoming forms coexist in variation in the speech community. However, the evidence also suggests that change is not an inevitable outcome of variation; certain sociolinguistic variables, such as the *–in/–ing* alternation in English (cf. alternations like *running* ~ *runnin'*), appear to have existed for many centuries without one form completely supplanting the other. But this asymmetry between the two is not unexpected in an adequate dynamic model of language. In expanding our view of grammar to incorporate variability, we do not preclude stability; synchronically some features of language do not vary, and diachronically, some features of language do not change, at least within certain time horizons. Hence diachronically stable variation is a possible characteristic of an adequate model of language.

What has conventionally been treated as two topics – 'linguistic variation' and 'language change' – is thus really one topic differentiated only by time scale: change is long-term variation. Consequently, each of these topics illuminates the other. Studies of variation in a short time frame (i.e. 'synchronically') implicitly contain information about what, from the perspective of a longer time frame (i.e. 'diachronically'), may be seen as change. This prospect has inspired a great deal of work within sociolinguistics and variation studies, addressing issues such as:

- what does linguistic variation today tell us about recent and future change? (e.g. how can change be read off from the synchronic record of diversity?)
- what does the study of language change tell us about variation today? (e.g. how does knowledge about change influence the interpretation of synchronic variation?)
- how does change proceed and progress?
- how can variation and change studies address traditional questions of historical linguistics?
- how does the social embedding of variation play out in diachrony?

This chapter presents a survey of contemporary issues in the study of variation and change, along with a reflection on the relationship of this work to the traditional approaches to language change embodied in the field of historical linguistics. We conclude with a consideration of the implications of this work for linguistic theory.

8.2 The study of change in progress

The earliest work on change in progress, by Labov (1963, 1966), made a basic distinction between two types of change that differed according to their social and psychological properties, what Labov called change from above and change from below. In this model, changes from above are effected consciously (hence Labov's elaboration as 'above the level of conscious awareness') and involve imitations of external models, while 'changes from below' are 'below the level of conscious awareness', and involve spontaneous innovations that are not based on an external model. Subsequent work confirms the need to recognise distinct types of change, based on different social mechanisms, but the specific criterion of consciousness is of doubtful utility in making the distinction, since there are changes involving spontaneous innovations of which there is considerable conscious awareness (e.g. the spread of high rising terminal intonations in declaratives in Australian English, Guy *et al.* 1986), and changes involving accommodations to external models that speakers show little awareness of. Instead, the literature suggests a convergence on a three-way distinction between **spontaneous** innovations, arising from within the speech community (subsuming 'change from below'), **borrowings** involving language or dialect contact but conducted by native speakers of the variety undergoing change (including Labov's 'change from above'), and **impositions**, arising in contact situations but conducted by speakers involved in language shift (cf. Thomason and Kaufman 1988; Van Coetsem 1988; Guy 1990). This third type has no equivalent in Labov's dichotomy; it includes the transferences from L1 to L2 that underlie 'foreign accents' and substratum effects.

Since each of these types involves a distinct social and psychological mechanism, it is expected that they display different social and linguistic distributions. Impositions should, at least initially, reflect systematic features of the L1, including patterns of variation; to the extent that progress in acquisition of the target language tends to suppress transference, one might expect gradual convergence on the variable patterns of the L2, led by speakers who have greater access to the target. Changes of the borrowing type most typically involve borrowing of prestige norms from a source external to the speech community, meaning that they are led by higher status speakers, and speakers in the age and class groups that have greater mobility, investment in, and/or access to the external prestige norm. Thus Labov (1966) finds that the reintroduction of coda /r/ in New York City is led by the upper middle class, and by young adults rather than adolescents. But note that other motivations for borrowing exist, and imply other social distributions; thus Cutler 2002 describes the adoption of features from African American English by white youth affiliated with hip-hop music and culture, and Stuart-Smith *et al.* (2007) describe the spread of historically non-local features like TH-fronting (i.e. replacement by /f/) among working-class Glasgow speakers, apparently motivated by the construction of a distinctive identity differentiating them from 'posh' (i.e. middle-class) speakers.

Spontaneous innovations do not involve contact with external sources; hence their social distribution reflects internal social dynamics of the speech community. Since they diverge from existing usage, rather than converging on a target, they reflect social processes of differentiation – contrastive processes of identity formation in which groups or individuals, by advancing the change, distinguish themselves linguistically from some reference point, rather than accommodating to it.

8.2.1 The social distribution of change in progress

Much of the research on the social distribution of change in progress has focused on spontaneous innovation ('change from below'). More than forty years of research has revealed a set of social tendencies that are well validated, at least for the types of societies in which these studies have been done – which, admittedly, are predominantly advanced industrial societies in the western world, although there are studies from Latin America (e.g. Cedergren 1973 on Panama, and numerous studies of Brazilian Portuguese), from Japan (e.g. Hibiya 1996), Egypt (Haeri 1996a), Iran (Modaressi 1978), and elsewhere.

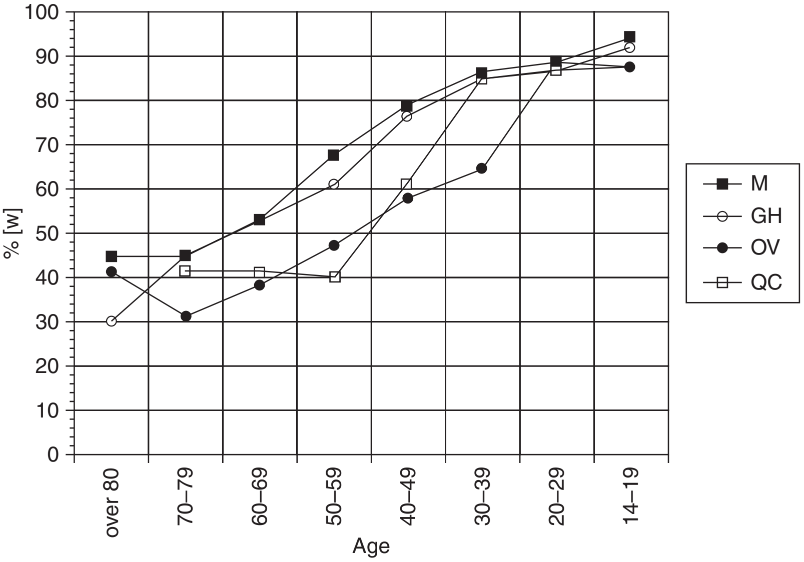

Figure 8.1 Percentage of speakers with [w] not [hw] in words like *which* and *whine* in four Canadian regions: Montreal (M), Golden Horseshoe (GH), Ottawa Valley (OV), Quebec City (QC) (from Chambers 2002: 63)

This work identifies age, class, and gender as the principal social dimensions reflecting ongoing spontaneous linguistic change in a community. The most commonly observed empirical patterns are as follows.

Age: The synchronic age distribution of a variable is considered the most crucial evidence for spontaneous change in progress. Older speakers are conservative, while younger speakers lead in the use of an innovation. The age distribution typically follows an S-shaped curve, as seen in Figure 8.1, from Chambers' research on the loss of /h/ in /hw/ clusters in Canadian English (Chambers 1998, 2002).

Such a distribution, recurring in many studies, is the typical synchronic face of ongoing change. It should be noted that where information is available on younger age-groups (children and younger adolescents), the data almost invariably show that the highest rate of innovation is not found among the youngest speakers, but rather among older adolescents and young adults, that is, somewhere in the age range 15–24. This no doubt reflects the development of social autonomy and the formation of a distinct social identity; the youngest speakers live in their parents' homes, and lead lives strongly governed by adults who, according to the age distribution seen in Figure 8.1, are relatively conservative. It is only in late adolescence or young adulthood that speakers construct an independent social and linguistic identity, achieve social autonomy, and minimise parental linguistic influence. This, evidently, is the point in the lifespan when speakers advance the use of linguistic innovations, going beyond what the next older age cohort has done to a still higher level of usage.

Social class: The social-class distribution of a spontaneous innovation has been argued by Labov and others also to display a distinctive pattern which is absent in cases of social stratification without ongoing change. This position is not entirely uncontroversial, but there are numerous studies reflecting the distribution that Labov considers decisive: the so-called 'curvilinear pattern', in which the peak use of an innovation is found towards the middle, or lower middle, of a social spectrum, while both the lowest and highest status groups lag in adopting the innovative form. Some classic examples are found in Figure 8.2, showing the distribution of vocalic changes in Philadelphia English. In both figures, the most advanced forms (in these cases, those raised the farthest along the front vowel diagonal) are found in the upper working class, as defined by a composite scale of socioeconomic status based on measures of occupation, education, and income.

The social motivation for the curvilinear pattern has been much debated. Labov's (2001) explanation relates the phenomenon to the differential importance of 'local identity' (i.e. solidarity with one's friends, family, neighbours, and community) across social classes. This is low both in highest-status groups (cf. concepts like the 'jet-set', people who are not strongly tied to one place, but derive their social position from supralocal affluence and influence), and lowest-status groups (cf. groups like the homeless, who also lack strong ties to a specific neighbourhood). For Labov, this aspect of social and psychological identity peaks in the upper working and lower middle classes, who have strong community ties and relatively low mobility. Hence these are the people with the greatest motivation to adopt and extend the distinctive characteristics of the communities that they belong to, and to demonstrate community membership contrastively by differentiating themselves from other individuals who do not belong. This view is reinforced by Milroy's work, showing strongest use of local forms by speakers with the strongest local community ties (1987).

Gender: Studies of the gender distribution of spontaneous innovations are distinctly skewed: substantially more of them show female speakers in the lead. This topic has attracted a great deal of interest in the field. But there are some studies showing males in the lead (e.g. the centralisation of the nucleus of /ay/ (in PRICE words) before voiceless segments in Philadelphia), and studies with no significant gender differentiation. The empirical findings are thus more mixed than those for age and class, and the explanations that have been proposed are more diverse.

Some of the major lines of explanation that have been advanced for gender differences in change are as follows. One approach refers to networking and socialisation patterns. Labov 2001 finds that the leaders of change are people,

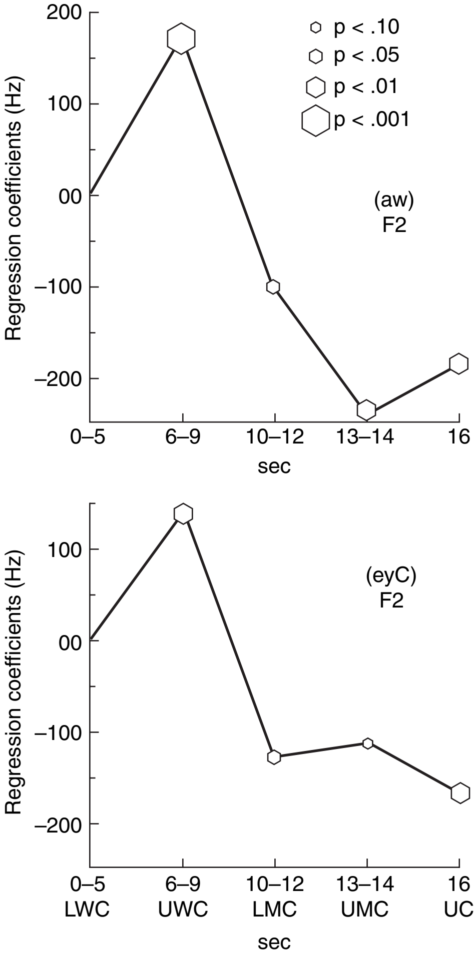

Figure 8.2 Curvilinear socioeconomic class distribution of vowel changes in Philadelphia English (from Labov 1980)

often women, who have both strong local ties and broader networks. In communities where women are more socially connected with a range of interlocutors, they would be more likely to have access to innovations as they develop and advance, and to participate in the construction of a local community identity via language. Another explanation appeals to gender differentiation in contact with younger children: if most of the adult caregivers for young children in a community are female (mothers, childcare workers, primary school teachers, etc.), then gender differentiated innovations favoured by females are more likely to be transmitted to the next generation of language acquirers (cf. Labov 2001). Male-led changes face a transmission problem if men don't talk much to the children.

An interesting account of sound changes in terms of acoustic differences between men and women is found in Haeri 1996b. Haeri suggests that acoustic iconicity is involved arising from the size differences between men and women. Vowel-like sounds are acoustically defined by their formants, which are resonances of the vocal tract. Like all resonances, formants vary with the size of the resonating space (compare the high notes of a little piccolo with the deep bass notes of a big tuba). Hence larger speakers (like adult males) have lower formant frequencies, and smaller speakers (like adult women, and more extremely, children) have higher formants. In the front–back dimension of the vowel space, which is acoustically signalled by the second formant (F2), a higher formant value means that a female speaker's vowels sound relatively more fronted. In normal speech perception, hearers compensate for this difference by 'normalising': interpreting the formant values with reference to the apparent size of the speaker's vocal tract. But hearers presumably retain access to the raw acoustic difference at some level. In changes on the front–back dimension, hearers might then systematically interpret female productions as being marginally fronter. Surveying nineteen different changes involving this dimension, Haeri notes that twelve out of thirteen fronting changes reveal a female lead, while five of six changes involving backing show males in the lead.

8.2.2 The linguistic distribution of innovations

Another productive area of research considers the question of directionality of linguistic change: do certain changes proceed only in one direction, and never reverse? If so, they permit the direction of change to be read off from the mere fact of variation. For example, grammaticalisations are changes that involve content words evolving into function words; the English indefinite articles *a/an* developed from the word *one*, for example. In Brazilian Portuguese *a gente*, historically a noun phrase meaning 'the people', is currently becoming a pronoun meaning 'we' (Zilles 2005). The reverse direction of change, of function word into content word, is virtually unknown. Hence, in the Brazilian case, the fact that content-word and function-word usages of *a gente* were in variation until recently (and may still be varying for some speakers) immediately implies that the content form is prior, and the pronominal usage is the innovation. In this particular case, a written historical record exists in which we can trace this development, but if grammaticalisation is unidirectional, the conclusion would still be valid even without any data from earlier times.

There are a number of claims in the diachronic linguistic literature that particular changes are unidirectional. As we have noted, grammaticalisations appear to be irreversible, perhaps because they typically involve a cluster of related changes in morphosyntax, phonology and meaning leading to the evolution of function words and affixes, which would be impossible to unpick. In phonology, cases of deletion are perhaps the most obvious candidate for unidirectionality; hence it is extremely likely that any cases of variation between the presence and absence of some element derive historically from the full form, via deletion, rather than the zero form, via insertion (barring cases of excrescent insertion such as Spanish *homre* > *hombre*, English *thunre* > *thundre* > *thunder*). Hence when we look at English final coronal stop deletion, encountering variable realisations of words like *just~jus'*, *old~ol'*, etc., we can be confident that *just*, *old*, etc. are closer to the historical sources. Other phonological processes that are much more common in one direction than the reverse include lenition, assimilation, and merger. Fortition (the reverse of lenition) is typically limited to specific prosodic conditions, when it occurs at all, and dissimilations are rare. Complete mergers are essentially irreversible, which is why they tend to spread across the dialectological and sociolinguistic landscape; however, there are attested cases of near-merger in which speakers retain some capacity to distinguish the merged phonemes, which occasionally leads to subsequent re-differentiation (see Thomas, this volume, also Labov 1994).

One well-known case where unidirectionality has been claimed is the theory of vocalic chain-shifts advanced by Labov, Yaeger and Steiner 1972 (LYS; also Labov 1994; for further discussion of vowel shifts, see Thomas, this volume). Based on extensive empirical studies, these scholars propose three principles governing chain shifts in vowel systems (such as the English Great Vowel Shift):

I. Tense vowels raise.
II. Lax vowels fall.
III. Back vowels move to the front.

Tense vowels, for these scholars, are those that relatively peripheral, articulated near the perimeter of the vowel space. An example of the first two of these principles is the English Great Vowel Shift. The non-high long vowels of Middle English (tense and peripheral) were all raised – for example, ME [eː] and [oː] raised to Modern English [iː] and [uː] respectively – while the high long vowels of ME first diphthongised and acquired centralised (i.e. non-peripheral and therefore lax) nuclei, which then fell down the central vowel space, so that ME [iː, uː] yield ModE [ay, aw].

These unidirectional principles are illuminating when we examine cases involving vocalic variability. Canadian English, for example, has systematic variation in realisation of the lax front vowels [ɪ, ɛ, æ], each of which varies along a range from higher and fronter to lower and backer (Clarke *et al*. 1995; De Decker 2002). LYS's principle II predicts that the direction of the change is towards the lower realisations, and indeed, other evidence, such as the age distribution of the variants in the population, confirms this prediction. This Canadian Shift involves lax vowels lowering in a chain shift, possibly triggered by the prior merger of /a/ and /oh/ (i.e. the vowel classes of COT and CAUGHT), which created room in the low central region of the vowel space for /æ/ to lower and back, generating a pull-chain shift.

In striking contrast to the Canadian Shift, the front lax vowels of English in Australia and New Zealand are raising, which appears at first blush to contradict LYS principle II. However, upon closer inspection, it turns out that these vowels are the most peripheral front vowels in these dialects. The front 'long' vowels (/ey / and /i/ as in FACE and FLEECE), have acquired centralised, non-peripheral nuclei in antipodean English, thereby abandoning the periphery to /ɪ, ɛ/, which are raising following LYS Principle I.

8.3 Real and apparent time

The underlying unity of linguistic variation and change is perhaps clearest in the analysis of the temporal extension of linguistic variables. There are two traditional perspectives on this question. One viewpoint has been described in the previous section: we can examine the age distribution of a variable at one point in time. This is customarily referred to as **apparent time** evidence, and it typically shows that the innovation is used more by younger speakers (Guy *et al.* 1986; Bailey *et al.* 1991). The alternative is akin to the perspective of traditional historical linguistics: examining data from different points in time, to see how the usage of variants has shifted during the interval between the samples. Such an approach looks at evidence from **real time**; for all innovations that are continuing to advance, real-time evidence will show an increase in the occurrence of the newer forms across time. (In language, as in genetics, not all changes are successful; some innovations appear, advance and then recede. Thus Blake and Josey (2003) found that the vocalic innovations – centralisation of the nuclei of /ay, aw/ (in PRICE and MOUTH words) – that Labov had described in Martha's Vineyard in 1963 were disappearing forty years later, as the island economy was re-oriented away from fishing toward more integration with the mainland.)

The relationship between these two kinds of evidence, real and apparent time, has received much attention. The two basic findings – spontaneous innovations show greater use by younger speakers in apparent time, and greater occurrence at later points in real time, suggest an obvious social mechanism for the spread of linguistic change. The community is not changing as a whole – with every speaker moving in the same direction, but rather, the membership of the community is changing, as new generations arrive and older ones depart, and different generations speak differently. The time course of a change spreading through a community thus involves two separate principles. First is incrementation: in a linguistic change that advances continually to completion, each successive age cohort uses on average a higher frequency of the new form than the cohorts that preceded it (their older siblings, in effect). Second is individual constancy: each cohort, and for the most part, each individual, remains mostly stable in their usage once they reach some point in late adolescence or young adulthood.

Real-time evidence supplements apparent-time evidence in an important way: it rules out an alternative hypothesis that might explain apparent-time data by a different mechanism, namely, age-grading – a situation in which individuals regularly alter their behaviour as they get older, but the community is not changing. We have interpreted Figure 8.1 as showing that [hw] clusters in *which, whine*, etc. are being lost in Canadian English. But why do we not believe that every Canadian speaker starts out using primarily [w] in these words when they are young, and proceeds gradually across their life-span to prefer more [hw] usage, peaking in old age? This is a logical alternative, which would not imply any change in the community as a whole; rather, it would imply that if we repeated Chambers' study at multi-year intervals, we would get essentially the same graph for age distribution, but each individual, were we able to track them, would have increased their [hw] usage as they aged. There are various reasons to be dubious of such an interpretation, including our understanding of normal language acquisition, but there is much real-time evidence in the literature that convincingly refutes such explanations. An example appears in Hibiya's (1996) study of the change of the velar nasal to [g] in Tokyo Japanese, shown in Figure 8.3.

This graph combines real-time and apparent-time data. The individuals to the left of the vertical line in Figure 8.3 are speakers interviewed by Hibiya in 1986. They follow a standard S-curve with younger speakers using more [g]. The speakers to the right of the line are people recorded by Japanese national radio in the 1940s who were born in the late nineteenth century. They are plotted according to the age they would have been in 1986 when Hibiya interviewed the other speakers, and they show the extension of the lower end of the curve into the past. But when they were interviewed, their actual ages were sixty–seventy-five years; a comparison of these subjects with Hibiya's subjects of comparable ages recorded forty years later rules out the possibility that all speakers start with high [g] use and decline as they get older. Rather, it is birth year and generational cohort that is associated with rate of [g] use, not age at any given point in time.

There are two approaches to collecting real-time evidence that permit the most detailed picture of a change in progress, and maximise the comparability of data across time. These are **panel studies** and **trend studies**. A panel study follows a specific group of individuals and resamples them at various points in

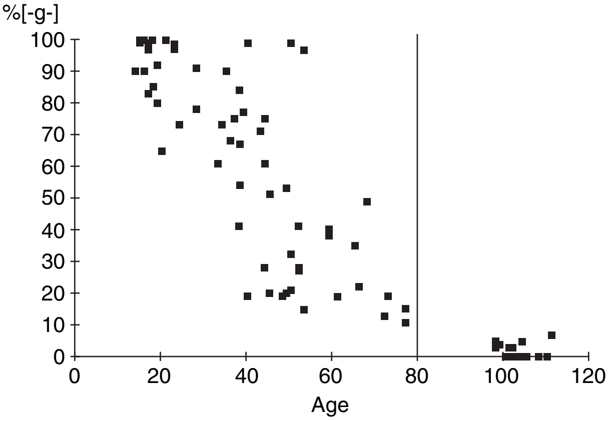

Figure 8.3 Apparent-time and real-time data: denasalisation of the velar nasal in Tokyo Japanese (from Hibiya 1996)

time. This makes it possible to test whether individuals indeed remain constant in their usage of variables across time. However, a panel study does not show us what subsequent generations do, to examine the progress of incrementation. The method that addresses this problem is the trend study, which examines successive cross-sections of the population at different points in real time. With reasonable comparability of the successive samples, especially with respect to the major social dimensions involved in linguistic change (age, class, and gender), a good trend study provides a moving picture of the change in progress, showing the generational advance of an innovative form.

Real-time studies, both trend and panel, have recently been a major area of research in language variation, as a result of the maturity of the field. Although it is rare to encounter a study in linguistics that is planned in advance to last for decades, what many linguists have done is to opportunistically replicate earlier research that indicated change in progress, in order to examine what has occurred in the community after the passage of a decade or more. In some cases this has involved new speaker samples, yielding a trend study, and in others, some original subjects have been recontacted, yielding a panel study.

One of the earliest (and still best) panel studies in sociolinguistic research builds on the Montreal French corpus, initiated in 1971 by D. Sankoff and G. Sankoff. Sixty speakers from the original sample were recontacted in 1984 by Thibault and Vincent (1990). Sankoff and Blondeau (2007) analysed a panel of thirty-two speakers who were recorded in both 1971 and 1984 for the use of the /r/ variable, which has been undergoing a change in Quebec French from apical to dorsal pronunciations. Of particular interest are the data in Figure 8.4, which shows the personal trajectories with respect

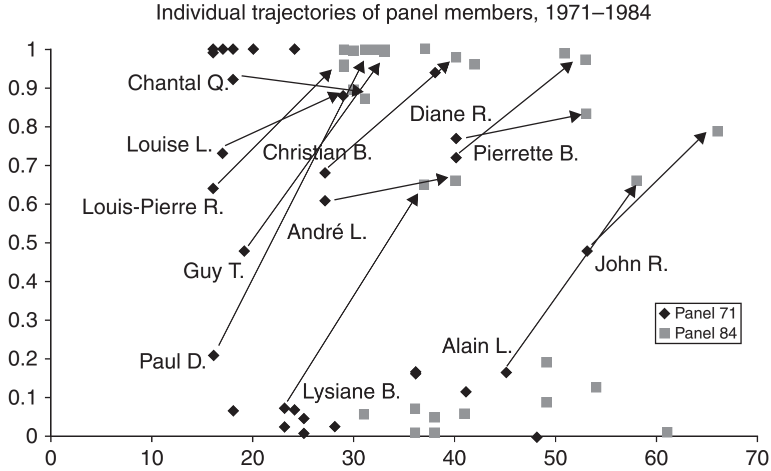

Figure 8.4 Real-time panel study of Montreal French speakers' use of [R] for /r/ (from Sankoff and Blondeau 2007).

to the change of thirty-two speakers across thirteen years. The figure shows that individual constancy is indeed the norm for speakers who were beyond the age of twenty-five when first recorded. Of speakers younger than this, those who markedly favoured one or the other variant when first recorded also show constancy of usage in later years (with one exception, Lysiane). The speakers who shift markedly in their usage are those who had intermediate rates of usage in the earlier sample, and who were young when first recorded. Only two significant exceptions to these generalisations are evident (Alain and John).

One striking consequence of the mechanism of change by cohort incrementation is that the S-curve of the age distribution of change in apparent time is also replicated in the real-time advance of change. Since trend studies with socially stratified samples are a recent methodological development, few of them exist with more than a few decades of time depth to illustrate this point, but it appears clearly in longer-term studies using written documents. One example is found in Kroch's work on the rise of English periphrastic *do* in questions and negative declaratives (1989a, 1989b, 2000). The modern form of this construction first appears in late Middle English in variation with earlier constructions involving subject–verb inversion in questions (e.g. *Do you eat fish?* varies with older inverted construction *Eat you fish?*) and postverbal negation in negative declaratives (*I don't eat fish ~ I eat not fish*). This was a spontaneous innovation in English which has no equivalent in any of the neighbouring languages with which English had contact in the Middle English period. After *do*-periphrasis

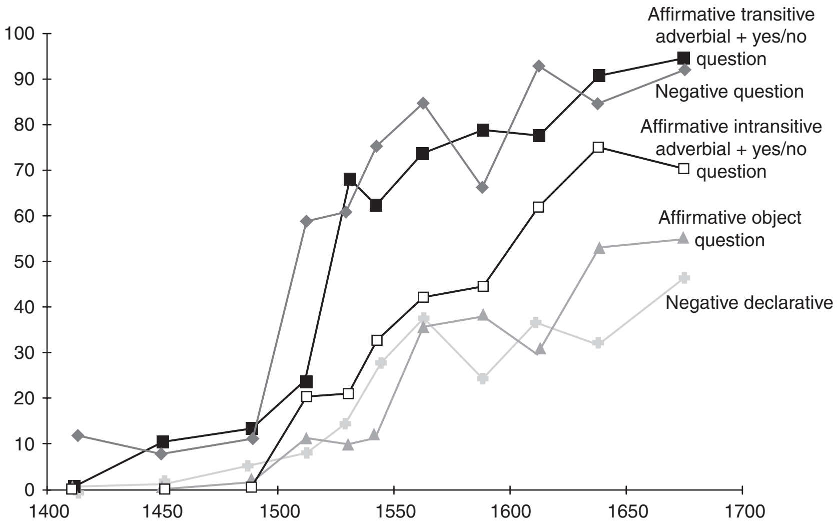

Figure 8.5 The rise of periphrastic *do* in Middle and Early Modern English (from Kroch 2000)

appears, it continues to vary with the older forms for about 400 years. Across that time span, the use of *do* rises along an S-shaped curve, as can be seen in Kroch's (2000) figure, reproduced here as Figure 8.5.

Five different contexts for the use of *do* are each plotted with a separate line on this graph. Although the data are somewhat noisy, causing a certain amount of jitter in the lines, each context progresses on a basic S-curve. Indeed, Kroch 1989b shows that these contexts are all statistically equivalent to S-curves defined by a logistic equation. The noisiness of the data is to be expected in a study based on historical documents, which do not afford us controlled and stratified samples of community usage. The original data supporting this figure were collected by Ellegård (1953) from documents and manuscripts whose date of provenance could be reasonably well established. But what is not often known for such data is the author's sociolinguistic identity – age, sex, class, dialect background, and residential history, for example – dimensions along which usage of this innovation most likely varied. Since at any given point in real time, document writers in England included people who differed on all these dimensions, each data point in this graph is a partly random selection from a cloud of possible values that might have been obtained from other writers in the same community at the same time. Given generational incrementation of change in adolescence followed by individual constancy in adulthood, not knowing the age of the authors is a clear source of noise: if the source documents in 1525 were written by young people aged about twenty-five, and the documents from 1575 were the product of elderly people of about seventy-five, the curve would appear to be flat in this time period, because both samples were drawn from the same generation, who did not change their usage in the interim. Consequently, the curves in Figure 8.5 track shifts in the limits and tendencies of the change, rather than the values associated with some specific reference point or community mean. Since, as we have noted, changes sometimes reverse direction, it is possible that not every change follows a smooth S-curve anyway, but the available evidence indicates that this is the dominant temporal pattern.

The accumulated evidence combining real-and apparent-time data thus confirms that changes advance in a community by incrementation among adolescents and young adults, and that after this age, individuals mostly stabilise their usage. But there are still many open questions, especially those involving causation and motivation. Why increment and then stabilise? We have suggested that incrementation is associated with adolescent identity formation: an incoming innovation has a sociosymbolic value as new and youthful, and serves to differentiate the innovators from their elders. But then why do individuals stabilise their usage in adulthood? This is subject to various interpretations. It could have something to do with linguistic maturation, much as is argued for so-called 'critical-period' effects (the decline in the ability to achieve native-like competence in languages learned in adult life.) This account appeals to neuro-biological factors. But there is also a recognisable psycho-social element: individuals are purposeful agents, who always command a range of styles and registers, and always vary their usage, including use of innovations, for purposes of accommodation, contrastive differentiation, identity construction and performance, and so on. Hence it is also possible that individuals associate their rate of use of innovative forms with a generational identity they wish to preserve across their lifespan. This explanation would link age stratification of language with the age stratification that is evident in clothing and hairstyles, music preferences, personal adornment, and other social behaviours that express generational identity.

Recent work on language and identity draws attention to the complexity of the social meanings and motivations of linguistic innovations, and their use by individuals in identity construction (see Moore, this volume). Individuals make personal choices about the use of variables to show affiliations with groups, to express personal stances towards hearers or situations, and to reference social interpretations and evaluations that may be used to index identity traits. Linguistic innovations of all sorts provide rich material for this elaborate orchestration of personal identity; it is therefore perhaps remarkable that broad social trends of the kinds we have identified (e.g. the temporal S-curve, the curvilinear class pattern) consistently emerge.

8.4 Variation and change and historical linguistics

The deepening engagement of sociolinguistics and variation studies with language change has brought these fields into close contact with the traditional discipline of historical linguistics. Initially, these were complementary approaches to diachronic questions. Historical linguistics focused on real-time data, used written evidence (necessarily, since no sound recordings of speech existed prior to the late nineteenth century), and dealt with completed changes across large-scale time spans – change across centuries and millennia. The focus on completed changes typically implied invariant, categorical models and descriptions. Variationist studies introduced the use of apparent-time evidence, focused on speech, and dealt with changes in progress, over shorter time spans – decades and generations – using quantitative models and descriptions. But more recently, this neat division of labour and focus has eroded. Variationist analyses have been conducted of documentary evidence from times long past. Quantitative approaches have been brought to bear on completed long-term changes using written materials to study their time course and the variation that occurred while they were in progress. Sociolinguistic models of the mechanisms of change have illuminated historical questions. Whereas traditional historical linguistics sought, in effect, to use the past to explain the present, the addition of the variationist perspective to diachronic research has also, in Labov's words (1994: 9), made it possible to 'use the present to explain the past'.

Particularly noteworthy examples of the fruitfulness of this fusion of variation and diachrony have occurred in research on historical syntax, such as the previously mentioned work of Kroch on English, and other studies on languages as diverse as Yiddish (Santorini 1993), Greek (Ann Taylor 1994), and Portuguese (Tarallo 1996). Phonological change is less amenable to this kind of approach, because of the limitations of orthographic evidence; nevertheless, some fruitful work has been undertaken, such as Toon's study of 'the politics of early Old English sound change' (1983).

The principal impact of variation research on historical linguistics, however, may be less methodological and empirical, and more theoretical. The variationist perspective has had little impact on the comparative method and the reconstruction of proto-languages, but understanding that variation is an essential way station in the course of change has substantial implications for evaluating the plausibility of the changes that are postulated and their social settings. For example, change-in-progress studies show that the period of variation can last for a long time, and multiple changes may be underway simultaneously. This has implications for reconstructing the sequencing of events (e.g. chain shifts). Contemporary variables sometimes exhibit lexical conditioning or irregularity, with implications for the regularity of sound change. Studies of the sociolinguistic types of change have implications for evaluating prior language contact. The transmission of lexical items, for example, implies borrowing as a primary mechanism of language change; hence in England after the Norman conquest, the huge inventory of French loanwords in English implies that native speakers of English were borrowing from French, while the paucity of phonological and syntactic effects suggests that Norman accents had very little impact on the English of descendants of the conquerors who underwent language shift. By comparison, English in India, which shows phonological characteristics common in Indian languages, such as retroflex consonants, and syntactic phenomena such as an invariant tag question '*isn't it*', is a case in which the main mechanism of change was imposition.

A significant consequence of the variationist perspective in historical studies has been the development of quantified models of change. Yang (2001) is a noteworthy example of this trend. Treating syntactic change as the product of the interaction between the distribution of syntactic structures in the input and the choices that child language learners face in their construction of a mental grammar, Yang proposes a probabilistic model of grammar competition that drives change forward along an S-shaped time course. The model crucially depends on variation: child language learners do not construct a single, static, invariant grammar to account for all the facts they encounter; rather, they entertain multiple alternatives, and select among them probabilistically.

Another significant contribution of variationist studies to historical linguistics is the refined view they permit of the stages of change. Conventional historical studies, relying on reconstruction from fragmentary evidence, typically account only for the endpoints of a change; a diachronic statement like *x* > *y* tells us that an early form *x* is realised centuries later as *y*, but gives no perspective on what happens during the intervening years. But synchronic studies of changes in progress make it possible to investigate triggering events and onsets of change (the actuation phase), and subsequent expansion of the innovation (the implementation phase).

Actuation appears to involve both social and linguistic factors; thus Labov (2010) attributes the original generalised tensing of /æ/ in the Inland North of American English to a social event: the early nineteenth-century mixture in north central New York state of speakers coming from several different dialect regions (including New England and southern New York), during the construction of the Erie Canal. These source dialects had different contexts for /æ/ tensing. The new communities that emerged from this mixture koinéised these conflicting patterns by tensing /æ/ in all contexts. The completion of the Erie Canal provided the pathway to settlement of the Upper Midwest, disseminating the new vowel phonology across a wide area. The tensed /æ/ vowel subsequently raised, vacating the low front corner of the vowel space; this provided a linguistic trigger for the fronting of the other low vowels. The implementation of this change had far-reaching effects on the vowels of this region, following the linguistic principles of vocalic chain shifting discussed in section 8.2.2, and ultimately yielding the complex vowel rotation known as the Northern Cities Shift (LYS, also Labov, Ash and Boberg 2006).

8.5 Change and linguistic theory

Why does language change at all? Accounting for language change in linguistic theory is a long-standing problem in linguistics, dating from at least the Neogrammarians. As we have noted, Saussure famously denied the relevance of diachrony to synchronic linguistic theory, and his position has been widely emulated for a century. Nevertheless, linguists of all theoretical camps have been unhappy with this drastic division of the field, and many have attempted (even if uneasily) to model change within whatever theoretical framework they favoured. Thus in the structuralist framework, sound changes were characterised as involving phonemic mergers, allophonic splits, alterations in phonetic values, and the like. (cf. Hoenigswald 1960). In the generative period, changes were described in terms of rule additions, losses, reorderings, and so on (cf. King 1969). Recent theoretical developments such as optimality theory (e.g. Anttila 1997) and exemplar theory (e.g. Bybee 2001) have often sought to explicitly incorporate accounts of linguistic change within their models.

Optimality theory has proven to be an exceptionally flexible framework for modelling change. The theory postulates a universal inventory of constraints, each stating some desirable phonological state of affairs; where languages differ is merely in the hierarchical rankings of these constraints (plus, of course, differing lexical inventories). Since any change in ranking defines a different grammar, and a different potential 'language', both variation and change can be subsumed into the OT account of language typology. Variable realisation of a final consonant, such as final –s and –r deletion in Caribbean Spanish and Brazilian Portuguese, and final –t deletion in English and Dutch, are modelled as variable rankings of constraints that militate against syllabic codas and those that favour faithful surface realisations of underlying segments. When faithfulness constraints are more highly ranked, the segment surfaces, but when the 'no coda' or 'simple coda' constraints prevail, surface realisations without the final segments are preferred. This is the typological difference between languages with open syllables (e.g. Yoruba), and those with closed syllables (English, Spanish, etc.), so the same mechanism can be pressed into service to account for variation and change. Variation is modelled by postulating variable ordering between the relevant constraints, and change across time is modelled by postulating a diachronic reordering of the relevant constraints (see, for example, Anttila 1997, 2002a, 2002b; Kiparsky to appear).

Exemplar theory is a recent development that places variation and change at the core of the model, relying on the naturally occurring variation in the input as the driving force (cf. Bybee 2001; Pierrehumbert 2002). This theory eschews abstract representations, postulating instead that speakers remember, in rich phonetic detail, the tokens of words that they hear pronounced, or produce themselves. Therefore, speakers have memories of the full range of variants they have encountered, and use these memories (the ‘exemplar cloud’) as targets for their own production, which then necessarily varies as well. The theory emphasises natural phonetic processes such as lenition and assimilation as the driving force in phonological change; words that are often repeated are more subject to these processes, altering the exemplar clouds of speakers in the direction of the change produced by the process. This model emphasises the importance of lexical identity and lexical frequency in variation and change, predicting that lexical items may differ (i.e. lexical diffusion), and that frequent words should lead sound change. Lexical diffusion has long been advocated in historical linguistics as an alternative to the exceptionless sound change model of the Neogrammarians (see Wang 1977; Labov 1981; Phillips 2006), based on a number of empirical cases where lexical irregularities are found in historical changes. Exemplar theory provides a formal model to account for such facts.

The most widely used theoretical framework in studies of variation and change, growing out of Weinreich, Labov and Herzog 1968, is the ‘variable rule’ (VR) model, a broadly generativist model in which optional elements in a grammar are probabilistically quantified (see Cedergren and Sankoff 1974 and Sankoff 1978 for further discussion). This is the dominant model in variation studies, and its extension to modelling change is straightforward, but has subtle and substantive implications.

The VR model postulates that any variable process may be subject to two conceptually different quantitative forces. First are contextual conditions: most variables have a lumpy distribution across the language, occurring often in some context and rarely in others: tensed and raised variants of /æ/ in English dialects, for example, are more common in pre-nasal contexts, and rarer or less advanced before voiceless stops. Deletion of final –t, and –d in English is more frequent before a following word beginning with a consonant, and rare before a following vowel. But second, there are also overall differences in the rate of use of any given variant, in different speakers, social-class groupings, speech styles, age cohorts, and so on. A particular dialect may have tensed /æ/ more frequently or more advanced phonetically than another dialect, even though both favour tensing in the pre-nasal context. A working-class speaker may delete /–t,–d/ more than a middle-class speaker, even while both delete more before consonants than before vowels.

This distinction between overall rate of use and contextual effects is captured in the VR model by two kinds of factors. Each process is associated with an ‘input’ probability, or p_0, which captures overall rate of use. In addition, a process may be associated with multiple contextual constraints, capturing the quantitative effects of favouring and disfavouring environments that promote or retard the selection of a particular variant. These are factor weights or partial probabilities associated with contexts, p_i, p_j, p_k, etc.

Given this model of variation, what changes across time is typically the overall rate of use of the innovative variant. Just as speakers and social groups differ in overall use while preserving the same constraint effects, and vary their overall rate in different speech styles while leaving contextual effects unaltered, successive age cohorts across the course of a change will increment the overall rate of use, leaving context effects unchanged. Change is change in the value of p_0, while the constraints on variable selection (p_i, p_j, p_k …) do not change.

The constancy of contextual effects across time has been demonstrated in a number of empirical studies, beginning with the work by Kroch illustrated above in Figure 8.5. Kroch formulates this observation as his ‘constant rate hypothesis’ – the claim that the rate of change in all contexts is the same. Kroch shows that the rate rises in English periphrastic *do* in all the contexts investigated in the figure are mathematically equivalent; that is, the logistic transform of each of the curves is a straight line with an essentially identical slope. Therefore, the most plausible interpretation is not that each context represents a separate change proceeding at an independent pace, but rather that there is only one change, following a single time course, governed in variable rule terms by a single p_0.

Syntactically, this single change can be described as a loss of V-to-I (verb to INFL) raising; in old and early Middle English, in sentences without auxiliaries, a main verb could move up to the high position (a.k.a. INFL) in the clause an auxiliary would occupy, thus preceding a negative, for example, (e.g. *They know not what they do*, with main verb *know* preceding *not*, parallel to *They must not know*, where auxiliary *must* precedes *not*). In Modern English, however, a main verb cannot occupy that position; instead *do* is inserted as a dummy auxiliary just when the main verb becomes separated from that position by some other material, such as a negative (You INFL not eat fish → *You* do *not eat fish*), or an inverted subject (INFL you eat fish? → Do *you eat fish?*’). The several contexts of the change show differences in their intrinsic favourability to the innovative variant that are stable across time. Each successive age cohort across the 400 years of the change was less and less likely to permit V-to-I raising, triggering the alternative solution of *do*-periphrasis at progressively higher rates.

The constant rate hypothesis follows directly from the VR model, distinguishing overall rates of use from contextual effects. Indeed, the constancy of the rate of change across different contexts constitutes important evidence in favour of VR for both variation and change. Alternative models of change, such as the OT treatment involving constraint re-ranking, lack any overall parameter comparable to p_0. This implies that change in an OT model should not be a smooth S-curve, but rather, a step function with inconstant contextual rates: each time a pair of constraints is re-ranked, the contexts they affect should show abrupt changes in the rate of occurrence of the variant realisations, while unaffected contexts would show no change. This is at odds with the empirical evidence.

8.6 Conclusion

Work on language variation has, since its earliest inception, addressed questions of language change. Nearly fifty years of research on these problems has turned up a substantial body of knowledge demonstrating that variation and change are in essence a single phenomenon viewed from different perspectives. This discovery requires linguists to develop new methodologies and theoretical approaches that make possible an integrated understanding of what the orthodoxy of twentieth-century linguistics treated as belonging to opposed and unrelated synchrony and diachrony. This is part of a broader integrative trend in twenty-first-century linguistics, bringing the insights of many disciplines together to tackle big issues that they were unable to resolve separately. At the centre of this integration are questions about stability and dynamism in language: why do languages change, and why do they (sometimes) remain the same? These are the questions that research on variation and change is helping to answer.

8.7 Where next?

Readers interested in following up this topic with further study would do well to examine the three volumes of Labov's *Principles of Linguistic Change* (1994, 2001, 2010), which provides an extended treatment of many of these issues. Condensed discussions of the social distribution of changes in progress may be found in Labov 1980 and Guy *et al*. 1986. The sociolinguistic typology of change is treated at length in Thomason and Kaufman 1988 and Van Coetsem 1988, and more succinctly in Guy 1990. Eckert 2000 is a classic source for the relationship of variation and change to social identity. A thorough discussion of the question of regularity in sound change is found in Labov 1981. Sankoff and Blondeau 2007 provide an excellent discussion of the relationship of real and apparent time evidence.