1 *The rôle of history*

1.1 Internal and external evidence

Any linguist asked to provide candidate items for inclusion in a list of the slipperiest and most variably definable twentieth-century linguistic terms, would probably be able to supply several without much prompting. Often the lists would overlap (*simplicity* and *naturalness* would be reasonable prospects), but we would each have our own idiosyncratic selection. My own nominees are *internal* and *external evidence*.

In twentieth-century linguistics, types of data and of argument have moved around from one of these categories to the other relatively freely: but we can identify a general tendency for more and more types of evidence to be labelled *external*, a label to be translated 'subordinate to internal evidence' or, in many cases, 'safe to ignore'. Thus, Labov (1978) quotes Kuryłowicz as arguing that historical linguistics should concern itself only with the linguistic system before and after a change, paying no attention to such peripheral concerns as dialect geography, phonetics, sociolinguistics, and psycholinguistics. Furthermore, in much Standard Generative Phonology, historical evidence finds itself externalised (along with 'performance factors' such as speech errors and dialect variation), making distribution and alternation, frequently determined by introspection, the sole constituents of internal evidence, and thus virtually the sole object of enquiry. In sum, 'If we study the various restrictions imposed on linguistics since Saussure, we see more and more data being excluded in a passionate concern for what linguistics is *not*' (Labov 1978: 275–6).

Labov accepts that 'recent linguistics has been dominated by the drive for an autonomous discipline based on purely internal argument', but does not consider this a particularly fruitful development, arguing that 'the most notorious mysteries of linguistic change remain untouched by such abstract operations and become even more obscure' (1978: 277). He consequently pleads for a rapprochement of synchronic and diachronic

study, showing that advances in phonetics and sociolinguistics, which have illuminated many aspects of change in progress, can equally explain completed changes, provided that we accept the uniformitarian principle: 'that is, the forces which operated to produce the historical record are the same as those which can be seen operating today' (Labov 1978: 281). An alliance of phonetics, sociolinguistics, dialectology and formal model-building with historical linguistics is, in Labov's view, the most promising way towards understanding the linguistic past. We must first understand the present as fully as possible: 'only when we are thoroughly at home in that everyday world, can we expect to be at home in the past' (1978: 308).

Labov is not, of course, alone in his conviction that the present can inform us about the past. His own approach can be traced to Weinreich, Labov and Herzog's (1968: 100) emphasis on 'orderly heterogeneity' in language, a reaction to over-idealisation of the synchronic system and the exclusion of crucial variation data. However, integration of the synchronic and diachronic approaches was also a desideratum of Prague School linguistics, as expressed notably by Vachek (1966, 1976, 1983). Vachek uses the term 'external evidence' (1972) to refer solely to the rôle of language contact and sociocultural factors in language change; this work has informed and influenced both contact linguistics and Labovian sociolinguistics. Although Vachek accepts external causation of certain changes, however, he still regards the strongest explanations as internal, involving the language's own structure. This leads to attempts to limit external explanation, often via circular and ultimately unfalsifiable statements like Vachek's contention (1972: 222) that 'a language system ... does *not* submit to such external influence as would be incompatible with its structural needs and wants'. For a critique of the internal/external dichotomy in this context, see Dorian (1993), and Farrar (1996).

More relevant to our discussion here is Vachek's argument that synchrony is never truly static: 'any language system has, besides its solid central core, its periphery, which need not be in complete accordance with the laws and tendencies governing its central core' (1966: 27). Peripheral elements are those entering or leaving the system, and it is vital that they should be identified, as they can illuminate trends and changes in the system which would not otherwise be explicable, or even observable. Peripheral phonemes, for instance, might be those perceived as foreign; or have a low functional yield; or be distributionally restricted, like English /h/ or /ŋ/ (Vachek 1976: 178). A dynamic

approach is therefore essential: the synchronically peripheral status of certain elements allows us to understand and perhaps predict diachronic developments, while the changes which have produced this peripherality can in turn explain irregularities in the synchronic pattern. This is not to say that Vachek collapses the two; on the contrary, his review of Chomsky and Halle (1968) is particularly critical of 'the lack of a clear dividing line that should be drawn between synchrony and diachrony' (1976: 307). Vachek considers Chomsky and Halle's extension of the Vowel Shift Rule from peripheral, learned forms like *serene* ~ *serenity*, to non-alternating, core forms like *meal*, an unjustified confusion of synchrony and diachrony: by in effect equating sound changes and synchronic phonological rules, Standard Generative Phonology in practice significantly reduces the useful conclusions which can be drawn about either.

Although Vachek seems to regard synchronic and diachronic data and analysis as mutually informing, the relationship is seen rather differently in Bailey's time-based or developmental linguistics. Bailey (1982: 154) agrees that 'any step towards getting rid of the compartmentalization of linguistics into disparate and incompatible synchronic, diachronic, and comparative or dialectal pursuits must … be welcomed', and proposes polylectal systems sensitive to diachronic data. He coins the term 'yroëth' (which is theory spelled backwards) for 'something claiming to be a *theory* which may have a notation and terminology but fails to achieve any deep-level explanation … All synchronic–idiolectal analysis is yroëthian, since deep explanation and prediction are possible only by investigating and understanding how structures and other phenomena have developed into what they have become' (Bailey 1996: 378). It is therefore scarcely surprising that Bailey regards the influence of diachronic on synchronic analysis as one-way, arguing that historical linguists are fundamentally misguided in adopting synchronic frameworks and notions for diachronic work: in doing so, they are guilty of analysing out the variation and dynamism central to language change by following the 'nausea principle': 'if movement makes the mandarins seasick, tie up the ship and pretend it is part of the pier and is not meant to sail anywhere' (Bailey 1982: 152).

We therefore have four twentieth-century viewpoints. The standard line of argumentation focuses on synchrony; historical evidence here is external, and is usable only as in Chomsky and Halle (1968), where sound changes appear minimally recast as synchronic phonological rules.

Vachek, conversely, argues that synchronic and diachronic phonology are equally valid and equally necessary for explanation. Labov argues that the present can tell us about the past, and Bailey the reverse. My own view is closest to Vachek's: if we are really to integrate synchrony and diachrony, the connection should cut both ways. That is, the linguistic past should be able to help us understand and model the linguistic present: since historical changes have repercussions on systems, an analysis of a synchronic system might sometimes benefit from a knowledge of its development. Perplexing synchronic phenomena might even become transparent in the light of history. But in addition, a framework originally intended for synchronic analysis will be more credible if it can provide enlightening accounts of sound change, and crucially model the transition from sound change to phonological rule without simply collapsing the two categories.

This book is thus intended as a contribution to the debate on the types of evidence which are relevant in the formulation and testing of phonological models, and has as one of its aims the discussion and eventual rehabilitation of external evidence. There will be particular emphasis on historical data and arguments; but issues of variation, which recent sociolinguistic work has confirmed as a prerequisite for many changes (Milroy and Milroy 1985; Milroy 1992), will also figure, and some attention will also be devoted to the phonetic motivation for sound changes and phonological rules.

However, although these arguments are of general relevance to phonologists, they are addressed here specifically from the perspective of one phonological model, namely Lexical Phonology. In short, the book also constitutes an attempt to constrain the theory of Lexical Phonology, and to demonstrate that the resulting model can provide an illuminating analysis of problematic aspects of the synchronic phonology of Modern English, as well as being consistent with external evidence from a number of areas, including diachronic developments and dialect differences. I shall focus on three areas of the phonology in which the unenviable legacy of Standard Generative Phonology, as enshrined in Chomsky and Halle (1968; henceforth SPE) seriously compromises the validity of its successor, Lexical Phonology: these are the synchronic problem of abstractness; the differentiation of dialects; and the relationship of sound changes and phonological rules. I shall show that a rigorous application of the principles and constraints inherent in Lexical Phonology permits an enlightening account of these areas, and a demonstration that

generative models need not necessarily be subject to the failings and infelicities of their predecessor. Finally, just as the data discussed here are drawn from the synchronic and diachronic domains, so the constraints operative in Lexical Phonology will be shown to have both synchronic and diachronic dimensions and consequences.

1.2 Lexical Phonology and its predecessor

Lexical Phonology (LP) is a generative, derivational model: at its core lies a set of underlying representations of morphemes, which are converted to their surface forms by passing through a series of phonological rules. It follows that LP has inherited many of the assumptions and much of the machinery of Standard Generative Phonology (SGP; see Chomsky and Halle 1968). LP therefore does not form part of the current vogue for monostratal, declarative, non-derivational phonologies (see Durand and Katamba 1995, Roca (ed.) 1997a), nor is it strictly a result of the recent move towards non-linear phonological analyses, with their emphasis on representations rather than rules (see Goldsmith 1990, and the papers in Goldsmith (ed.) 1995). Although elements of metrical and autosegmental notation can readily be incorporated into LP (Giegerich 1986, Pulleyblank 1986), its innovations have not primarily been in the area of phonological representation, but rather in the organisational domain.

The main organisational claim of LP is that the phonological rules are split between two components. Some processes, which correspond broadly to SGP morphophonemic rules, operate within the lexicon, where they are interspersed with morphological rules. In its origins, and in the version assumed here, the theory is therefore crucially integrationist (but see Hargus and Kaisse (eds.) 1993 for discussion, and Halle and Vergnaud 1987 for an alternative view). The remainder apply in a postlexical, postsyntactic component incorporating allophonic and phrase-level operations. Lexical and postlexical rules display distinct clusters of properties, and are subject to different sets of constraints.

As a model attempting to integrate phonology and morphology, LP is informed by developments in both these areas. Its major morphological input stems from the introduction of the lexicalist hypothesis by Chomsky (1970), which initiated the re-establishment of morphology as a separate subdiscipline and a general expansion of the lexicon. On the phonological side, the primary input to LP is the abstractness controversy. Since the

advent of generative phonology, a certain tension has existed between the desire for maximally elegant analyses capturing the greatest possible number of generalisations, and the often unfounded claims such analyses make concerning the relationships native speakers perceive among words of their language. The immensely powerful machinery of SGP, aiming only to produce the simplest overall phonology, created highly abstract analyses. Numerous attempts at constraining SGP were made (e.g. Kiparsky 1973), but these were never more than partially successful. Combating abstractness provided a second motivation for LP, and is also a major theme of this book.

The problem is that the SPE model aimed only to provide a maximally simple and general phonological description. If the capturing of as many generalisations as possible is seen as paramount, and if synchronic phonology is an autonomous discipline, then, the argument goes, internal, synchronic data should be accorded primacy in constructing synchronic derivations. And purely internal, synchronic data favour abstract analyses since these apparently capture more generalisations, for instance in the extension of rules like Vowel Shift in English from alternating to non-alternating forms. However, as Lass and Anderson (1975: 232) observe, 'it just might be the case that generalizations achieved by extraparadigmatic extension are specious'; free rides, for instance, 'may just be a property of the model, rather than of the reality that it purports to be a model of. If this should turn out to be so, then any "reward" given by the theory for the discovery of "optimal" grammars in this sense would be vacuous.' In contrast, I assume that if LP is a sound and explanatory theory, its predictions must consistently account for, and be supported by, external evidence, including diachronic data; the facts of related dialects; speech errors; and speaker judgements, either direct or as reflected in the results of psycholinguistic tests. This coheres with Churma's (1985: 106) view that '"external" … data … must be brought to bear on phonological issues, unless we are willing to adopt a "hocus pocus" approach … to linguistic analyses, whereby the only real basis for choice among analyses is an essentially esthetic one' (and note here Anderson's (1992: 346) stricture that 'it is important not to let one's aesthetics interfere with the appreciation of fact'). The over-reliance of SGP on purely internal evidence reduces the scope for its validation, and detracts from its psychological reality, if we accept that 'linguistic theory … is committed to accounting for evidence from all sources. The greater the range of the evidence types that a theory is capable of handling

satisfactorily, the greater the likelihood of its being a ''true'' theory' (Mohanan 1986: 185).

These ideals are unlikely to be achieved until proponents of LP have the courage to reject tenets and mechanisms of SGP which are at odds with the anti-abstractness aims of lexicalism. For instance, although Mohanan (1982, 1986) is keen to stress the relevance of external evidence, he is forced to admit (1986: 185) that his own version of the theory is based almost uniquely on internal data. Elegance, maximal generality and economy are still considered, not as useful initial heuristics, but as paramount in determining the adequacy of phonological analyses (see Kiparsky 1982, Mohanan 1986, and especially Halle and Mohanan 1985). The tension between these relics of the SPE model and the constraints of LP is at its clearest in Halle and Mohanan (1985), the most detailed lexicalist formulation of English segmental phonology currently available. The Halle–Mohanan model, which will be the focus of much criticism in the chapters below, represents a return to the abstract underlying representations and complex derivations first advocated by Chomsky and Halle. Both the model itself, with its proliferation of lexical levels and random interspersal of cyclic and non-cyclic strata, and the analyses it produces, involving free rides, minor rules and the full apparatus of SPE phonology, are unconstrained.

Despite this setback, I do not believe that we need either reject derivational phonology outright, or accept that any rule-based phonology must inevitably suffer from the theoretical afflictions of SGP. We have a third choice; we can re-examine problems which proved insoluble in SGP, to see whether they may be more tractable in LP. However, the successful application of this strategy requires that we should not simply state the principles and constraints of LP, but must rigorously apply them. And we must be ready to accept the result as the legitimate output of such a constrained phonology, although it may look profoundly different from the phonological ideal bequeathed to us by the expectations of SGP.

In this book, then, I shall examine the performance of LP in three areas of phonological theory which were mishandled in SGP: abstractness; the differentiation of related dialects; and the relationship of synchronic phonological rules and diachronic sound changes. If LP, suitably revised and constrained, cannot cope with these areas adequately, it must be rejected. If, however, insightful solutions can be provided, LP will no longer be open to many of the criticisms levelled at

SGP, and will emerge as a partially validated phonological theory and a promising locus for further research.

The three issues are very clearly connected; let us begin with the most general, abstractness. SGP assumes centrally that the native speaker will construct the simplest possible grammar to account for the primary linguistic data he or she receives, and that the linguist's grammar should mirror the speaker's grammar. The generative evaluation measure for grammars therefore concentrates on relative simplicity, where simplicity subsumes notions of economy and generality. Thus, a phonological rule is more highly valued, and contributes less to the overall complexity of the grammar, if it operates in a large number of forms and is exceptionless.

This drive for simplicity and generality meant exceptions were rarely acknowledged in SGP; instead, they were removed from the scope of the relevant rule, either by altering their underlying representations, or by applying some 'lay-by' rule and a later readjustment process. Rules which might be well motivated in alternating forms were also extended to non-alternating words, which again have their underlying forms altered and are given a 'free ride' through the rule. By employing strategies like these, a rule like Trisyllabic Laxing in English could be made applicable not only to forms like *divinity* (~ *divine*) and *declarative* (~ *declare*), but also to *camera* and *enemy*; these would have initial tense vowels in their underlying representations, with Trisyllabic Laxing providing the required surface lax vowels. Likewise, an exceptional form like *nightingale* is not marked [−Trisyllabic Laxing], but is instead stored as /nɪxtVngǣl/; the voiceless velar fricative is later lost, with compensatory lengthening of the preceding vowel, to give the required tense vowel on the surface.

The problem is that the distance of underlying representations from surface forms in SGP is controlled only by the simplicity metric – which positively encourages abstractness. Furthermore, there is no linguistically significant reference point midway between the underlying and surface levels, due to the SGP rejection of the phonemic level. Consequently, as Kiparsky (1982: 34) says, SGP underlying representations 'will be at *least* as abstract as the classical phonemic level. But they will be more abstract whenever, and to whatever extent, the simplicity of the system requires it.' This potentially excessive distance of underliers from surface forms raises questions of learnability, since it is unclear how a child might acquire the appropriate underlying representation for a non-alternating form.

A further, and related, charge is that of historical recapitulation: Crothers (1971) accepts that maximally general rules reveal patterns in linguistic structure, but argues that these generalisations are non-synchronic. If we rely solely on internal evidence and on vague notions of simplicity and elegance to evaluate proposed descriptions, we are in effect performing internal reconstruction of the type used to infer an earlier, unattested stage of a language from synchronic data. Thus, Lightner (1971) relates *heart* to *cardiac* and *father* to *paternal* by reconstructing Grimm's Law (albeit perhaps not wholly seriously), while Chomsky and Halle's account of the *divine* ~ *divinity* and *serene* ~ *serenity* alternations involves the historical Great Vowel Shift (minimally altered and relabelled as the Vowel Shift Rule) and the dubious assertion that native speakers of Modern English internalise the Middle English vowel system. I am advocating that historical factors should be taken into account in the construction and evaluation of phonological models; but the mere equation of historical sound changes and synchronic phonological rules is not the way to go about it.

Here we confront our second question: how are sound changes integrated into the synchronic grammar to become phonological rules? In historical SGP (Halle 1962, Postal 1968, King 1969), it is assumed that a sound change, once implemented, is inserted as a phonological rule at the end of the native speaker's rule system; it moves gradually higher in the grammar as subsequent changes become the final rule. This process of rule addition, or innovation, is the main mechanism for introducing the results of change into the synchronic grammar: although there are occasional cases of rule loss or rule inversion (Vennemann 1972), SGP is an essentially static model. The assumption is that underlying representations will generally remain the same across time, while a cross-section of the synchronic rule system will approximately match the history of the language: as Halle (1962: 66) says, 'the order of rules established by purely synchronic considerations – i.e., simplicity – will mirror properly the relative chronology of the rules'. Thus, a sound change and the synchronic rule it is converted to will tend to be identical (or at least very markedly similar), and the 'highest' rules in the grammar will usually correspond to the oldest changes. SGP certainly provides no means of incorporating recent discoveries on sound change in progress, such as the division of diffusing from non-diffusing changes (Labov 1981).

It is true that some limited provision is made in SGP for the restructuring of underlying representations, since it is assumed that

children will learn the optimal, or simplest, grammar. This may not be identical to the grammar of the previous generation: whereas adults may only add rules, the child may construct a simpler grammar without this rule but with its effects encoded in the underlying representations. However, this facility for restructuring is generally not fully exploited, and the effect on the underliers is in any case felt to be minimal; thus, Chomsky and Halle (1968: 49) can confidently state:

> It is a widely confirmed empirical fact that underlying representations are fairly resistant to historical change, which tends, by and large, to involve late phonetic rules. If this is true, then the same system of representation for underlying forms will be found over long stretches of space and time.

This evidence that underlying representations are seen in SGP as diachronically and diatopically static, is highly relevant to our third problem, the differentiation of dialects. The classical SGP approach to dialect relationships therefore rests on an assumption of identity: dialects of one language share the same underlying representations, with the differences resting in the form, ordering and/or inventory of their phonological rules (King 1969, Newton 1972). Different languages will additionally differ with respect to their underlying representations. The main controversy in generative dialectology relates to whether one of the dialects should supply underlying representations for the language as a whole, or whether these representations are intermediate or neutral between the realisations of the dialects. Thomas (1967: 190), in a study of Welsh, claims that 'basal forms are *dialectologically mixed*: their total set is not uniquely associated with any total set of occurring dialect forms'. Brown (1972), however, claims that considerations of simplicity compel her to derive southern dialect forms of Lumasaaba from northern ones.

This requirement of a common set of underlying forms is extremely problematic (see chapter 5 below). Perhaps most importantly, the definition of related dialects as sharing the same underlying forms, but of different languages as differing at this level, prevents us from seeing dialect and language variation as the continuum which sociolinguistic investigation has shown it to be. Furthermore, the family tree model of historical linguistics is based on the premise that dialects may diverge across time and become distinct languages, but this pattern is obscured by the contention that related dialects are not permitted to differ at the underlying level, while related languages characteristically do. It is not at all clear what conditions might sanction the sudden leap from a situation