The reductive character of articulatory evolution

Richard Mowrey & William Pagliuca

Although directionality in internal phonetic evolution has been noted by many historical linguists, current models of change have been unsuccessful in incorporating this property directly into their theoretical structures. This paper explores some of the issues standing in the way of such an enterprise, including the definition of which kinds of changes must be accounted for and the selection of descriptive primitives. A general theory of unidirectional reductive articulatory evolution is advanced which unites the shallow diachronic processes found in casual speech with longer-term processes common in the historical record*.

1. Background

Although synchronic concerns have held central position in linguistic description and theory-building for many decades, the study of the change of sound patterns over time, which initiated modern linguistics, has remained an integral part of the discipline, with prominent figures of each generation addressing sound change in major statements. Despite numerous advances in the level of sophistication with which sound patterns have been analyzed, however, there is still no consensus among historical linguists on the answers to fundamental questions concerning sound change.

Prominent among the unresolved questions are those having to do with the very nature of sound change. Is sound change a phenomenon *sui generis*, distinct from such things as borrowing and analogy, and thus deserving of separate theoretical and analytical treatment? Is it essentially a physical (phonetic) or psychological (phonological) phenomenon? If phonetic in nature, is its source to be sought in production or perception? Are sound changes phonetically abrupt or gradual? Are they lexically abrupt or gradual? What is the relation between synchronic variation and change?

The sorts of answers offered to such questions are in part determined by the prevailing views of synchrony and diachrony at the time they are addressed. Even within one generation of linguists, however, a wide variety of emphases, as well as contrarian views, are typically evident. Although not every statement on sound change articulates a position on all issues, in the last few decades advocates for each of the

Rivista di Linguistica, 7.1 (1995), pp. 37-124 *(ricevuto nell'ottobre 1994)*

possible answers to many of these questions can be found. A necessarily incomplete survey of some of these positions follows.

'Mentalist' claims notwithstanding, the generative tradition continues what was begun in earnest by structuralist phonemics, in which the primacy of the role of psychological over physical units in both synchrony and diachrony is presupposed.[1] Phonologists may routinely allude to the phonetic basis of such relatively well-understood assimilatory phenomena as vowel nasalization, but phonetic modes of explanation have not been incorporated into phonological representations. They rather have the status of an unformalized codicil, invoked when the need arises, but otherwise regarded as having relatively little direct bearing on theory and description. Phonology is, after all, primarily about distinctiveness, and this, rather than the phonetic character of lexical and grammatical material, is what its machinery is designed to represent. For generativists, sound change is grammar change (Kiparsky 1965, Postal 1968); since grammar change takes place in minds, sound change (or that aspect of it which such a theory regards as worthy of being addressed) cannot be physical.

Among those who do not subscribe to the prevailing, phonological, view of sound change is Ohala, who in numerous publications (1971, 1974a-b, 1975, 1978, 1981, 1983a-b, 1990, among others) has championed a physicalist approach to sound change, along the way revealing flaws in certain types of phonological reasoning. For Ohala, sound changes arise as the result of production and auditory perception constraints, in which the role of the listener is crucial, though generally underappreciated. Thus he maintains that although 'much of the synchronic variation in speech – from which diachronic variation arises – can be traced to the speaker or the physical principles which map articulation to sound [...]', variation is 'due to "innocent" misapprehensions about the interpretation of the speech signal or [...] to listeners' inexperience' (Ohala 1990:266). For this reason, though he has identified changes in the relative timing of articulations as responsible for certain sound changes, his focus has tended to be on the role of the listener, and hence on perceptual factors in driving change. We return to a consideration of Ohala's position at the end of this section.

That misperception and/or misproduction are implicated in the initiation of sound changes is also assumed in certain phonological approaches to change. Bailey (1977) seems to believe in a widespread propensity for misperception and other error, though his general theoretical stance is that change is directional in the sense that it moves structures from more to less marked (Bailey 1973). Andersen (1972, 1973) has interpreted a number of changes as innovations arising as

principled resolutions of perceptually problematic sequences. Limits on the faithfulness and accuracy of the perceptual and motor systems for speech also figure in Hockett's (1965) conception of sound change. Expanding on an account in Jespersen (1922), Hockett argues that speech sounds are properly characterized in terms of ranges of pronunciation ('acoustic allophones') around targets or norms. Sound change is then the 'drift' of these acoustic allophones, 'and hence of distinctive features', over time:

> Sound change is not reducible to analogy: the latter leads only to the replacement of one array of distinctive features by another as the 'realization' of some lexical or grammatical form. Sound change is affected by certain kinds of borrowing, in that so-called 'fashions of pronunciation', as in imitation of a prestigious model, can alter the density distribution; but it is not REDUCIBLE to borrowing because the density distribution is largely altered by innumerable tiny imprecisions of pronunciation and by constant channel noise – the kind of thing Jespersen was talking about, but mostly on a much finer-grained scale – that takes place totally out of awareness. (p. 202)

> Sound change CAN go on because language is redundant: most of the time, a hearer need register only a suitably scattered small fraction of the stigmata of the speech signal in order to know what has been said. Language, in turn, is redundant because otherwise it would have had no survival value for our species. Sound change DOES go on because of NOISE. Only in a permanently noiseless universe would sound change cease [...] (p. 203f.).

The role of redundancy in enabling sound change and changes in morphological and syntactic constructions over time is clear. That channel noise has the relation to change that Hockett claims for it, however, seems problematic. Channel noise and variation about norms, whether due to imprecise production or not, together predict an inherently directionless drift, whereas many common sound changes are unidirectional. We expect, for example, aspirated stops to reduce to fricatives and fricatives to [h], and vowel + nasal consonant sequences to have nasalized vowels as descendants, but the reverse directions of drift – [h]'s becoming fricatives becoming stops, nasalized vowels evolving into oral vowel + nasal stop sequences – seem rare. Nor can channel noise explain the register-dependent variation in individual speakers reflected in the formal-to-casual speech continuum. Here too we find clear indications of directionality: as we move from the more formal to the more casual end of the continuum, we observe increasing lenition and assimilation, not a random mix of lenition and assimilation, fortition and

dissimilation. If over time sounds do not 'wander' (p. 202), but rather move in particular directions, something other than channel noise must be responsible.

Well-attested changes have repeatedly taught us what to expect in terms of both the direction and magnitude of individual sound changes. Given correspondence sets $p^h : p^f : f : h : \varnothing$ and $an : \tilde{a}$, no one hesitates in choosing p^h and an as the original forms (cf. Lass 1978). Similarly, no one, given a synchronic alternation of [pʰ] and [h] with a pattern of distribution most economically describable as h → pʰ would seriously regard such a rule as a faithful reflection of the diachronic events which gave rise to the distribution. The difficulty is that such expectations are nowhere formalized, but are rather part of a storehouse of extratheoretical and extranotational knowledge. Our expectations concerning magnitude fare no better. Although changes with well-attested histories provide guidelines – given a correspondence $p^h : h$, we do not assume that there were no intermediate stages – viable definitions of what constitutes a minimal or possible phonological change have proved elusive. The result is that periodically we need to be reminded, as by Andersen (1972), that a correspondence $A:B$, whether across contemporaneous dialects or temporally separated stages of the same language, is not the same as a diachronic event $A > B$. Precisely because correspondences are sometimes all we have access to, it would seem more than reasonable to incorporate what we know of direction and magnitude in framing a theory of the kinds of events sound change can and cannot give rise to.

Kiparsky (1965) articulates a generativist position, arguing for the psychological over physical nature of sound change, for acquisition as its locus, and for its abrupt rather than gradual character. The first two points follow from viewing language as grammar change[2] and from the parent theory's framing the general problem in terms of acquisition. The choice of abruptness is motivated by metatheses and apparently articulatorily-unconditioned changes of place of articulation, which do not evidence the intermediate stages presupposed by a view of change as phonetically gradual.

Two differences between the approaches of Hockett and Kiparsky seem especially noteworthy. By acknowledging the 'basic triad' (p. 190) of borrowing, analogy, and sound change, Hockett allows each to be treated, in principle, as an independent phenomenon to be described and explained. Whether or not one agrees with the particular description or explanation offered, the separation of the phenomena in a given domain into classes based on known aetiological distinctions is a sound first step toward arriving at a realistic account of the entire domain. This is especially so if the classes evidence non-random distribution of particular kinds of diachronic events or if the events in each typically display

differences in rate and manner of implementation. For Kiparsky, all diachronic changes are rule changes, so all are represented in a uniform feature-and-rule notation; distinctions of origin (sound change vs. borrowing vs. analogy) are not provided for, and are de facto treated as irrelevant. Inasmuch as these distinctions are thus rendered theoretically and descriptively inaccessible, this appears to be a step backward. Secondly, for Hockett, at least some inherent variation is a given and bears a direct relation to change, whereas in the generative paradigm variation is a fact about performance, can therefore safely be ignored, and is thus also irrelevant.

Obviously, generativists were aware of the traditional distinctions outlined by Hockett; it was their theoretical and analytic utility that was denied. Postal (1968:283) went so far as to dismiss the very possibility that any systematic causal basis underlying sound change could be uncovered, asserting that 'There is no more reason for languages to change than there is for automobiles to add fins one year and remove them the next, for jackets to have three buttons one year and two the next, etc.', and that therefore 'the "causes" of sound change without language contact lie in the general tendency of human cultural products to undergo "nonfunctional" stylistic change.' The difficulty with this view is that nonfunctional stylistic change, like channel noise, cannot account for directionality. If change over time were capricious – the product of social whims – we would expect the historical record to consist of an essentially random mix of changes, rather than displaying the broad patterns of reduction and assimilation so familiar to historical linguists.

Wang (1969), whose evaluations of Hockett's and Postal's positions are similar to ours (he faults Hockett for failing to account for directionality and Postal for ignoring the considerable evidence for phonetic motivation in many sound changes) makes a compelling case for lexical diffusion, arguing that sound changes are lexically gradual but phonetically abrupt – the reverse of the Neogrammarian position, by which sound change is phonetically gradual but lexically abrupt. Lexical diffusion studies (Chen & Wang 1975, Barrack 1976, the papers in Wang, ed. 1977, Krishnamurti 1978 among others) provide systematic demonstrations of what was known for some time – that lexical items do not all change at once. Schuchardt noted that frequently-used words are affected earlier by a given sound change (1885:26); Zipf, who argued similarly, documented the correlation between frequency of use and word length in a number of languages (1929, 1932). Though Wang dismisses word frequency as a significant factor (p.15, footnote 12), more recent studies have since confirmed Schuchardt's observation for certain assimilatory and reductive phenomena (e.g. Fidelholz 1975 for vowel

reduction and Hooper 1976b for schwa deletion in English, and Phillips 1980 on the raising of OE *a* before nasals).[3]

Wang also rejects phonetic gradualness, citing, as types of sound change for which it cannot be correct, those in which, in X > Y, X and Y involve different articulators "between which there is no continuum"; metathesis; and "segment addition or deletion, since many sound types are either present or absent, but never present in gradient quantities" (p. 13). Here Wang seems to overstate the case for phonetic abruptness. As we will see later on, for some changes of place of articulation for which no physiological continuum is apparent, articulatory pathways in fact exist. Others are known to be the result of language contact, and as such cannot be construed as evidence for abruptness in internal change. Metathesis is more of a challenge, but not every alphabetic permutation is proof of abrupt disruption of original articulatory sequencing. In regular metathesis, it is not unusual to find that one of the segments involved in the apparent transposition is articulatorily minimal (e.g. the glottals [h] and [ʔ] or labialization). What appears to be an interchange of two entire segments may then be the result of a change in the timing of a very small number of muscular events – similar to the phenomena which, we argue below, underlie assimilatory processes. Temporal changes may also be responsible for the apparent transposition of segments whose component articulations are largely identical, as in the regular Semitic metathesis of the final [t] of the reflexive prefix with an immediately following sibilant (Malone 1971), as well as for other instances, such as the familiar dialectal variation in Eng. *iron, bird.*

Although not as easy to come by as it is in domains where the physiological continuum is obvious (as, e.g. in the gradient movement suggested for ongoing vowel quality changes in Northern Shift speakers (Labov, Yaeger & Steiner 1972)), evidence for gradience in segment addition or deletion phenomena is also available. Fourakis & Port (1986) show that the transitional (inserted) stops arising from [ns] and [ls] sequences in American English, as in *dense* and *false*, are significantly shorter than the underlying stops in *dents* and *faults*. Electromyographic evidence for the gradient character of insertions in speech errors (usually assumed to be characterized by all-or-none segmental intrusions and permutations) is presented in Mowrey & MacKay (1990). The gradient nature of both synchronic form and diachronic changes is also accessible to casual observation. For example, Boas noted that in the Nass River dialect of Tsimshian final [n], [m] and [l] are "articulated but not pronounced unless non-final in the sentence", and that "vowels are suppressed in a similar manner by being only indicated by the position of the mouth, without being articulated" (Boas 1911). Similarly, some speakers of Spanish assert that, for the characteristically weak

intervocalic [ð] in frequent words and affixes, they can feel their tongue move toward, but not quite attain, contact (cf. Menéndez Pidal's (1968:100) use of superscripts to schematize the progression for the past participle suffix *-ado* as -aðo > -a^{ð}o > -a$^{(ð)}$o (and, 'más vulgarmente', > -ao)). In these and other cases, consonants are clearly decaying to zero in a phonetically gradual fashion.

The general difficulty here is that the case for abruptness cannot be judged on the witness of orthography and broad transcription, since alphabetic units by their nature will always testify that change is abrupt. For relevant testimony we must turn to phonetic accounts, whether based on the informal observations of a Boas or Menéndez Pidal or on instrumental data. Although we expect abrupt substitution of one sound for another in cases of language contact, such finer-grained data as we possess point to the phonetic gradualness of internal changes.

A corollary of lexical gradualness, as Wang noted, is that changes may take many generations to run to completion; this results in variation across generations in the number of items affected by a given change. Labov and his collaborators have shown that ongoing changes are typically characterized by variation; variation is not 'free' but systematic, and variants typically have social value. Bailey (1973:34) has also argued that since variation is controlled, it cannot be written off as performance, but is rather necessarily part of what linguists must describe.

Labov (1981) attempts to reconcile lexical diffusion with the Neogrammarian notion of phonetically gradual but lexically unconditioned change. Like Wang, Labov regards metathesis and 'other discontinuous consonant shifts', such as changes in place of articulation, as phenomena for which it is 'only natural' to expect lexical diffusion. For the ongoing vowel shifts he and his coworkers have documented, however, he argues that there is phonetic gradualness and regularity, that is, no sensitivity to lexical identity. Rather, what regular phenomena display is sensitivity to fine phonetic conditioning. The contexts for the vowel changes, for example, include the identity of individual pre-vocalic and post-vocalic consonants. As he points out, such specificity of contextual conditioning is a close equivalent of progressive word-by-word expansion through possible environments: with ever finer measures of phonetic context, we approach the limiting case, wherein each lexical item constitutes a separate definable environment. To determine whether diffusion of a change through the lexicon is in fact distinct from its progression through the most narrowly-definable phonetic contexts, Labov proposes that homonyms serve as a test case: if lexical identity rather than phonetic context determines applicability, homonymic pairs might be expected to behave independently. The test he reports on

compares F2 values of the vowel nuclei for the homonymic pairs *too / two* and *know / no* in a speaker from Philadelphia, in which /u/ and /o/ are in movement in the vowel space in an ongoing shift. The difference is greater for the members of the first pair, but in neither case do the statistical tests used show a significant difference in the extent to which the members of each pair have advanced, leading Labov to conclude that the progression through phonetic contexts is distinct from progression through the candidate items in the lexicon. Although in principle homonyms are ideal candidates for such a determination to be made, it is not clear that these data allow a fair test. If discourse frequency is one of the factors determining which items participate first in sound changes, then only where homonyms differ significantly in frequency would we expect to find a difference. But here all four test items are of extremely high discourse frequency, suggesting that all might well participate early and more or less equally. Some indication that homonyms can display such differentials is given by Keller (1987: 130), who reports the results of a pilot study showing significantly greater reduction in syllable duration and articulatory gestures for the grammatical members of grammatical-lexical homonymic pairs in French. Thus *la* 'the' is more reduced than *las* 'weary', and *vers* 'toward' more reduced than *vers* 'worm'.

Labov's general conclusion is that some vowel shifts are usually free of (reported) lexical conditioning, others are typically lexically sensitive, and that changes of consonantal manner are phonetically gradual but lexically abrupt, whereas changes of consonantal place are phonetically abrupt but lexically gradual. By the latter division, lenition phenomena such as fricativization are gradual but not lexically sensitive, whereas the coarticulation changes involved in assimilation are abrupt and proceed in an item-by-item fashion. In support of his position, Labov argues that numerous variation studies of lenition phenomena in Spanish and Portuguese fail to evidence lexical diffusion, exhibiting instead "every sign of Neogrammarian regularity" (p. 302). But lenition in Spanish does not always display such lexical uniformity; intervocalic voiced fricatives, for example, are typically more reduced in frequent words such as *agua* 'water' and *lado* 'side' than they are in less heavily used items. Evidence for the lexical sensitivity of both lenition and assimilation phenomena is available for other languages as well. Phillips (1984), also arguing that Labov's division is not maintainable, lists the raising of OE /a/ to /o/ before nasals (treated in Phillips 1980), the devoicing of final stops in Old English, the devoicing stage of the High German Consonant Shift (as described in Barrack 1976), and other cases. More recently, Quesada (1988), in a cross-dialectal investigation of phonological processes in Quechua, presents abundant evidence for

the lexical sensitivity of voicing, affrication and spirantization of stops. For example, syllable-initially, Proto-Quechua uvular *q spirantized in some varieties, voiced in others, and, in dialects in which the changes are in progress, the degree of reduction – the extent to which spirantization or voicing has advanced – is shown to be correlated with the frequency of individual words and grammatical morphemes (pp. 76-124). The distinction Labov wished to draw thus does not appear to be maintainable. For changes of all major types there is some evidence for both phonetic and lexical gradualness. Since we expect environmental expansion in lexical diffusion as well, we might conclude that both lexical identity and phonetic environment are relevant in the spread of most if not all changes.

Perhaps less systematically and intensively studied than age-related and social variables, but nevertheless clear (and clearly related), is the variation reflected in the formal-to-casual speech continuum. To some extent the stylistic continuum in effect encapsulates recent generational history, and thus serves to make part of the temporal continuum accessible. Roughly speaking, the more formal end of the continuum may be taken to represent what in prior generations were more casual modes of speech, with current casual modes reflecting subsequent changes. But the correspondences appear to be global in nature, reflecting not only the progress of relatively obvious ongoing changes, but apparently idiosyncratic reductive changes to lexical and phrasal material as well. This suggests that viewing diachrony in terms of successive waves of discrete processes, each of which alters a particular subpart of the structure of individual lexical items until all candidate items have been reached, may fall short of providing a complete picture (see 2 below).

To summarize to this point, sufficient evidence exists to justify specific answers to most of the questions concerning the nature of sound change. First, the aetiological differences between sound change proper and other phenomena effecting the change of form over time are so obvious that the de facto abandonment of the distinction by generativists may be viewed as an aberration. As Hockett argued, in order to construct a theory of sound change, it is necessary to distinguish sound change from borrowing and other phenomena. Secondly, that sound changes proceed in a lexically gradual rather than abrupt fashion has been more than amply demonstrated by Wang and others; the role of relative frequency in diffusion also appears to have been adequately documented. Third, evidence for the phonetic gradualness of at least some processes assumed to be necessarily abrupt, though for obvious reasons more difficult to obtain and therefore less abundant, is sufficient to suggest that the case for it has much more merit than Wang assumed. Fourth, as Labov and others have demonstrated, variation is not only systematic but often reflects the action of changes in progress, and is

thus necessarily within the domain of phenomena which a theory of sound change must address.

By arguing in support of phonetic gradualness we were also obviously taking the position that sound change is essentially phonetic rather than phonological, physical rather than psychological. The only outstanding issue, then, is whether the source of sound change more likely lies in production or perception.

Some changes which have been argued to have acoustic or auditory bases appear, on closer inspection, to have articulatory bases. For example, Ohala (1974b) analyzed the vocalization of velarized [l] as in Cockney as an auditorily-based substitution of [w] for velarized [l], appealing to the closeness of the two sounds on an F_1-F_2 plot and to the occurrence of [w] for velarized [l] in children's speech as evidence. But young children also substitute [w] for initial [l] and various kinds of r, which are not so close to [w] on an F_1-F_2 plot. More importantly, the misperception/substitution analysis is called into question by the close transcriptions of Sivertsen (1960), which make it clear that the [w] assumed to be uniformly substituted for the velarized lateral occurs only if the preceding vowel is high and round; otherwise, the reflex of the lateral is a glide which is labial or not, and high or mid, depending on the vowel. This is precisely what one might expect of a decaying lateral coarticulating with the vocalic articulations in its immediate environment. In particular, there is no introduction of labial articulations where none was originally present (some details on this and similar changes are given in Pagliuca 1982 and Pagliuca & Mowrey 1987).

Changes of place of articulation in fricatives for which auditory or acoustic analyses have been proposed are similarly amenable to articulatory analysis. One is the change in Middle English by which the original velar fricatives in words such as *cough* and *laugh* became labiodentals. Although often cited as an example which shows that auditory factors play a role in at least some diachronic changes (Ladefoged 1971:44; Hyman 1973), this change, which we examine in 5.1, also appears to have a straightforward articulatory interpretation.

Evaluating the status of changes such as $k^w > p$ and $\theta > f$, which Ohala and others attribute to acoustic and auditory factors, is made difficult by the fact that historical attestations of such changes are not unambiguously instances of internal change. Nor do they appear to occur in contemporary changes in progress or in casual speech. Relative to a host of diachronic reduction phenomena, then, such changes appear to be properly regarded as infrequent in internal change. This is not to deny that cases of $\theta > f$ for example occur in bilingual and other contact milieux, in which auditory perceptual factors are clearly implicated. To be sure, some of the changes which Ohala cites as due to perceptual

factors are in fact frequent. Thus he includes, in his (1974b) list of auditorily-based changes, [s, ç, x] > [h] and voiceless stop > [ʔ]. But such changes, in which only or principally laryngeal remnants of original fricatives and stops survive, are in our view more likely to be instances of articulatory reduction (see 5.3.2) than substitutions attributable to misperception. Note that these changes are not bidirectional.

Ohala has also suggested that auditory constraints play a crucial role in assimilation: hearers may attribute acoustic effects induced by a conditioning articulation to inherent properties of the conditioned segment, i.e. change results from the auditory misparsing and reassignment of properties (Ohala 1981); a similar role for auditory constraints is proposed by Hombert, Ohala & Ewan (1979) for certain cases of tonogenesis. Note that in such analyses unintended articulatory properties (coarticulation resulting from overlap of lingual articulators, or, for tonogenesis, 'small, consonantally-induced F_0 perturbations on vowels' (Hombert, Ohala, & Ewan 1979:38)) are identified as playing an initiatory role, setting up the conditions for reassignment. In positing subtler and finer-grained sorts of interaction between articulation and perception, these analyses are much richer than those which invoke the acoustic similarity of two sounds as the cause of the substitution of one segment for another. The question is whether perceptual mediation of this sort is essential in order to explain these intermediate-term effects of coarticulation. In 5.1 and 5.2 we argue that perceptually-based reassignment may constitute an unnecessary analytic step, i.e. that coarticulation changes alone (subsequently perceived and reproduced faithfully) can account for such phenomena.

Acknowledging the lack of bidirectionality of diachronic events, Ohala has suggested, on the basis of experiments on confusions in the visual and vibrotactile perception of letters, that it may be explained by assuming 'incomplete perception': the perceptual system misses some property of the stimulus and reports or produces a substitute which has all the properties of the original minus the missed feature (Ohala 1983b:235f.). Note that if we take this to predict a step-wise diminution of features (perhaps all the way to ∅), it is in principle interpretable as equivalent to the incremental reduction of articulation over time. But we are still left with the problem of reconciling such an account with the manifest ability of the perceptual and production systems to accommodate levels of detail many orders finer than those Ohala refers to.

The existence of dialectal detail, and its successful continuance in generational transmission, suggests not only that speakers routinely command levels of phonetic detail much finer than that which our descriptive tools can accommodate, but also that audition and articulation must be very closely linked. Some indication of the faithfulness and

accuracy of the auditory perception system in allowing individuals to acquire the necessary detail is illustrated by the results of a study conducted by Labov on a supposed merger between /ay/ and /oy/ in Tillingham (Essex). In particular, the reliability of speakers' reports of what they perceive is called into question by Labov's (1974:298-307) demonstration that, although Tillingham speakers cannot always hear (i.e. label as different) /ay/ and /oy/ in isolated lexical pairs when recordings are presented to them, spectrograms reveal that they nevertheless consistently produce them as different.

> In these cases, speakers consistently make small differences in natural speech which maintain the identity of word classes, but they cannot accurately label these differences on conscious reflection, either in their own speech or in the speech of their close associates who speak the same dialect [...] From these results we infer the strong possibility that /ay/ and /oy/ have remained in close approximation in Tillingham for several hundred years, heard as the same, yet not the same in fact. (p. 306).

We suggest that though auditory perception is implicated in changes arising in language contact, in which the differently-tuned speech perception systems of the speakers of an adopting language come into play, misperception does not provide a plausible explanatory basis for internal change, since the analyses depend on attributing to the speech perception system a tendency toward confusion and approximation which it does not appear to exhibit. Articulatory processes by which production patterns gradually change over time (which changes have acoustic and therefore perceptual consequences) thus appear to provide a better account of internal change. The nature of variation along the formal-to-casual speech continuum corroborates this conclusion. Not only does the relation between more formal and more casual speech modes seem to be most naturally characterizable as the product of unidirectionally reductive changes, but speakers effortlessly control the entire range of production along the continuum. This suggests that the reductive changes arise from on-line articulatory phenomena whose effects are incremental and continuous rather than from a series of events with a basis in misperception or reassignment.

2. Preliminaries

Although the foregoing provides some indication of how we view sound change, it does not adequately prepare the ground for us to outline what we believe to be a reasonable approach to constructing a theory of sound change. The principal reason for this is that certain aspects of our

approach to change – in particular, our perspective on articulation – are at odds with the views of others. Consequently, we begin here by making explicit how our approach can be viewed as one possible natural outcome of some of the material we have just reviewed.

For a given sound to change in some items without changing in others suggests that what is actually changing are properties of the particular sound-meaning expressions (words or phrases), rather than sounds as elements of a system or sounds in generalized contextual frames. In order for a speaker (or learner, or model of either) to keep track of which /A/'s or [A]'s have changed and which have not, and therefore correctly produce, (or identify, or characterize) both changed and not-yet-changed items, information about individual word identities and their differential phonetic realizations must somehow be directly linked. Although this is an obvious conclusion, it leads us to ask in what sense, and to what extent, compositionality can be said to exist. Are the expressions actually (i.e. in other than an analytic sense) built up from the elements we identify, or are the elements instead epiphenomena, derivable over expressions, but not their 'building blocks'. In the latter case, the phonetic character of the expressions would not be dependent on building blocks in any production-relevant sense – the expressions would be more like independent holistic production routines. On such an interpretation, lexical diffusion and other evidence of the phonetic adventurousness of particular lexical and grammatical stretches would not constitute anomalous behavior.

Viewed from a wider perspective, changes to individual lexical items form only part of the domain of phonetic change. Change is also always and routinely taking place in larger expressions, most obviously in phrases subject to frequent repetition. Repetition is also the driving force in the evolution of grammatical markers from lexical constructions in grammaticalization and in the creation of lexical items from phrasal antecedents (Bybee et al. 1994:5, 8, 19f.; Haiman 1994). In these cases, with repeated recruitment of the specific constructions or phrasal units, the original material undergoes continuous semantic and phonetic reduction over time.

Repetition also seems to drive phonetic change in lexical items generally, which is why we are not surprised to find frequently-used words routinely in the forefront of diffusing sound changes (Schuchardt 1885:27). In general, words, though typically intermediate in size between grammatical morphemes and phrasal expressions, display a much slower rate of phonetic change, since most individual lexical verbs and nouns necessarily have lower incidences of use than grammatical markers, greetings or often-repeated expressions. Everywhere along the continuum, however, from phrasal constructions to affixes, the relationship between repetition and reduction holds.

Greetings, grammatical markers, and certain other frequent expressions tend to be phonetically idiosyncratic, displaying irregularly reduced form. This encourages us to view their phonetic specificities as stored, rather than as the results of travel through on-line reductive and assimilatory processes. For example, the greeting *How are you doing?* is never phonetically as full as the simple concatenation of the individual words it is nominally composed of, nor would we want to treat its most reduced variants as if they derive anew each time by on-line phenomena from the ostensible full form.[4] Moreover, some of the processes frequently-repeated material appears to undergo are themselves idiosyncratic relative to the general processes of a given dialect. For example, the virtually complete decay of the interdental fricative in *What's the matter?* ([wətsəˈmæDr]) is not mimicked in the nearly identical *What's the Matterhorn?* In *Why don't you* > [ˈwãjntʃə], although vowel nasalization, palatalization, and the reduction of the vowel in *you* are familiar enough as English phenomena, there is no general process in English reducing or deleting [d] between vowels. Individual lexical items display similar behavior. For example, the common reduction to loss of [t] in *let* in (1a) (giving [ˈlɛ̃mi]) doesn't extend to the [t] of *set* in (1b) (*[ˈsɛ̃mi]), nor is the loss of [v] in *give* in (2a) ([ˈgĩmi]) possible for *sieve* in (2b) (*[ˈsĩmi]), though these differentials could not be predicted from the full forms of *let* and *set* and *give* and *sieve*. In (3a), *going* reduces, in the dialects of the authors, to [gõjŋ] or [gɔ̃n], but *sewing* in (3b), which should behave the same way, does not become [sõjŋ] or [sɔ̃n]. Individual items also exhibit differential sensitivity to more familiar assimilatory phenomena. Palatalization across word boundaries in American English is clearly sensitive to the frequency with which a particular verb + noun collocation occurs, which is in turn to some extent dependent on the frequency of the component items. Thus, the stressed [ju] following *can't* in (4a) typically participates, but the identical syllable in (4b) would not. (Cooper & Paccia-Cooper (1980), examining similar examples, show experimentally that the incidence of palatalization across such boundaries depends on the frequency of the [j]-initial item.)

(1) a. Let me down, I want to walk.
b. Set me down, I want to walk.

(2) a. Give me a cup of flour, would you?
b. Sieve me a cup of flour, would you?

(3) a. I'm going this way.
b. I'm sewing this way.

(4) a. Why can't YOU do it?
b. Why can't EUNICE do it?

Grammaticalized and grammaticalizing constructions display similarly advanced and idiosyncratic patterns of lenition and assimilation. To continue with English examples, [ˈgɔ̃nə] and other reduced variants of the future marker *going to* are not possible with *going to* when it expresses movement in space; [ˈãɲɔ̃nə] or [ˈãjmɔ̃nə] are casual speech possibilities in (5a), but not in (5b). [wãnt # tu] reduces to [ˈwãnə] in (6a), but not (6b). As an auxiliary, but not as a main verb, *have* may reduce to [ə]; for both (7a) and (7b), when read so as to be intonationally nearly identical, [ˈwUDəv] is possible, but further reduction, to [ˈwUDə] or to [ˈUDəv] or [ˈUDə], can only occur in (7a). In all these cases (cf. also Keller's results, cited above), despite the phonetic identity of the full forms of grammatical and lexical pairs, the patterns of differential reduction are clear, with the grammatical forms displaying much greater ranges of eroded variants (a record, in effect, of successive repetition-induced coarticulation changes and other reductive events).

(5) a. I'm going to sing.
b. I'm going to the store.

(6) a. I don't want to complain.
b. I'm not wont to complain.

(7) a. Houses made of this would have burned very quickly.
b. Houses made of this wood have burned very quickly.

For individual lexical items, in addition to changes which appear to diffuse through the lexicon steadily over some generations, there are also what at first glance appear to be very 'slow' changes, in which differential sensitivity persists for much longer periods of time. In English, for example, unstressed medial vowels have been reducing to zero in a lexical-item specific fashion for centuries (cf. Jespersen 1909). Thus vowels have been lost in *captain (< capitain), napkin (< nappekin)* but not in *capital*; in *wondrous* and *disastrous (< wonderous, disasterous),* but not yet in *ponderous* or *thunderous.* For such items as *battery, chocolate, sophomore, grocery, history, mystery, blustery* and *robbery,* there is individual and dialectal variation toward eventual loss of the vowel, but structurally identical items such as *flattery, sorcery* and *snobbery* tend not to participate. If we recognize that on-line reductive effects are everpresent molders of phonetic form, we are not forced to view such items as either random exceptions or as the erratic results of the application of discrete processes which 'got started' in the past but somehow keep on getting derailed, never reaching the items in their proper domain.

In all the classes of cases just discussed, the role of repetition, and hence the use-dependent nature of synchronic variation and change, is clear. If, for all the collocations involving paired items we have just cited, we assume that the starting points (the underlying or stored phonological forms) are the same, the non-participant items would have to be prevented from undergoing the processes or events which alter the high-frequency items. Whatever the nature of the basic units we choose, lexical identity has to be coded for; otherwise, control of such systematically varied outputs would not be possible and our models would not allow all and only the participating items to reduce and assimilate.

If the phonetic character of all material on the continuum is subject to continuous and cumulative modification over time in a use-dependent rather than strictly phonetic context-dependent manner, can we continue to assume that lexical entries are composites of standardized building block elements which are uniformly altered by phonetic processes? Or do we recognize a continuum of realizations of what we identify as a given phone or sequence of phones, with different items displaying systematically distinct, though partially overlapping, ranges of variation under identical phonetic conditions. If this is what the core material of languages really looks like – if, in simplified terms, not all [v]'s in a given phonetic context are actually the same and not all [t # j] sequences palatalize to the same degree (or, diachronically: at the same rate) – then the assumption that lexical and grammatical morphemes are built up from standardized elements which are modifiable in a given context in a regular and predictable fashion falls short of providing for accurate descriptions.

A comprehensive theory of production must somehow address phonetically idiosyncratic material and differential behavior under phonetic processes. If we reject the option of some form of holistic and detailed phonetic storage of individual expressions, then some way of tagging items that undergo phonetic processes more readily, undergo them to greater extents, or undergo processes that items with ostensibly structurally identical subparts do not, must be specified. Even if such idiosyncrasy were limited to greetings and set phrases, this would be a non-trivial matter, but exiling such structures from consideration in a building block approach in order to protect the integrity of the system or the assumptions of uniformity and regularity is a practical possibility only if their numbers and the types of oddity involved are relatively circumscribed. But by taking such a step, we would be failing to account for phonetic change properly speaking, and instead be contenting ourselves with a sanitized phonological version of diachrony. If phonetic subparts of individual items in core lexicon can vary, subtly but systematically, from ostensibly identical subparts in other items, and if the extent to which individual items are susceptible to and altered by phonetic processes over time and synchronically is in part dependent on item-specific properties, something close to holistic storage must be viewed as a reasonable alternative.

The phonemic principle is assumed to be necessary for memory (Halle & Clements 1983; Lindblom et al. 1984; Halle 1985); holistic storage is not regarded as a possible alternative. The argument rests on the ostensibly prohibitive multiplicity of forms (words or root morphemes) which speakers must ultimately acquire and control. But given that the meaning and use of each of a large core set of forms must be learned in detail in order for an individual to acquire a language, we may question whether the argument from multiplicity really warrants the routine denial of an analogous specificity to phonetic form. We believe that it does not, and that the level of phonetic detail in the core lexical material of any given dialect in itself suggests that such assumptions cannot be correct.[5] In fact, given that speakers must acquire and store detail anyway, the linking of phonetic and semantic form – the pairing of sound and meaning that linguists regard as their responsibility to explicate – might be effected in a smoother way in a holistic model. Nor need this entail giving up any of the advantages accruing from the phonemic view of language, since it does not seem unreasonable to assert that most of what are referred to as phonemic and phonotactic facts are in principle reducible to the particular motor and perceptual 'tunings' of a given language or dialect.

We have argued that it is articulation directly – rather than articulation mediated by perception, or psychological units of the nature of phonemes – that changes in internal sound change, and that the locus of change is specific meaningful expressions, and not, or at least not in any primary sense, sounds or sound patterns independent of expressions. Use-dependent differentials in detail and in sensitivity to processes suggest that the acquisition and storage of phonetic properties of core lexical material is holistic in nature, as it appears to be in early childhood (Ferguson & Farwell 1975; Peters 1983; cf. also Studdert-Kennedy 1987). If so, it may be that much of what is learned and reproduced are particular and detailed phonetic-semantic routines. The sorts of things typically assumed to be the shapers and controllers of substance – inventories of basic units, phonotactic constraints, etc. – would then be epiphenomena, derived as multiple sets of generalizations over the production routines, and changed as the routines are altered by evolutionary events.

We now turn to a fuller treatment of what we regard as a reasonable approach to the construction of a theory of sound change.

3. Toward a general theory of sound change

The principal step in theory construction is not assembling observations on the phenomena to be accounted for, but classifying and thereby reducing the number of types of such observations. In constructing a theory of sound change, the most crucial classification is of the number of types of change – processes – accorded theoretical recognition.

A first approach to reducing the types of change to a smaller set will necessarily be statistical. It might be argued that in classifying sound changes into types, statistical inference has no place, and that all observations must be accorded equal status. Acceptance of this argument would not only isolate us from the rest of the sciences but would also forever relegate us to the role of cataloguing rule types with no real chance of explaining how language changes or why certain types of change are so often repeated in both related and unrelated languages.

Though current diachronic theory is mute with regard to the expected frequency of phenomena in which articulations are reduced (substantively, as in lenition and loss, and temporally, as in assimilation and other coarticulation changes) relative to those in which new articulatory material is introduced – epentheses and fortitions – instances of reduction vastly outnumber cases of augmentation. Though most historical linguists are aware of the great imbalance in the incidence of reductive vs. augmentative events over time, diachronists typically do not incorporate this observation into their theoretical structure.[6] Diachronic phonological theory, by being noncommittal towards this imbalance, thus tacitly regards augmentative and reductive events as equiprobable.

Diachronic data varies greatly in quality and completeness. The catalogue of changes may be viewed as including both attested and reconstructed changes, the latter suggested sometimes by comparative, sometimes by internal evidence, with allophonic and morphophonemic relations of the sort typically formalized in synchronic rule descriptions providing additional candidates for the status of "possible phonological change". Clearly, the corpus of reconstructed changes, whether considered as a whole or after classification into types, contains an unavoidable element of noise in the form of spurious individual events and types of change. Similarly, as is well known, synchronic phonological rules are not reliable indicators of the direction of diachronic events, since the most elegant solution to a set of data involving allophonic distribution or morphophonemic alternation may involve expressions which either telescope or reverse the events which unfolded over time (see e.g. Lass 1978). Since not all evidence is of equal quality, not all evidence deserves equal weight when one attempts to construct a theory of sound change.

When the quality of data is high, as it is for relatively recent and in-progress changes, what we seem to observe in internal change – change of phonetic form over time which is not precipitated by language contact or by analogical factors – is assimilation and lenition. Augmentative events, in which articulations not present in original forms are intruded, seem exceedingly rare.[7] In the most accessible high-quality data – the changes evident in casual/allegro relative to careful or lento speech – the absence of augmentative events is virtually absolute.

Putatively augmentative events such as the rise of transitional glides and stops have long been recognized to involve only change in the temporal relations of existing articulations rather than the intrusion of new articulatory material (see 6.1). Such events are thus fortitions or counterevidence to reduction only in notational systems incapable of describing their temporal character. Of greater moment is the fact that many examples of fortition and epenthesis are analytically-derived rather than observed – when not posited to have occurred in the remote past, they owe their existence to inferences drawn on the basis of comparative and internal reconstruction. Because phonetic records for much of the past are non-existent, because we are at the mercy of historical accidents – the survival of some dialects, the loss of others (cf. Lass 1978:250); literary monuments for some dialects, none for others – our reconstruction techniques must be informed by strong hypotheses derived from those observations of change which afford us the most complete and accurate phonetic records.

The great imbalance in the incidence of reductive versus augmentative events in language-internal changes generally, and the decrease in the incidence of observations of augmentative events as data quality increases, suggest the following hypothesis. Internal change of phonetic form over time is unidirectional, and the direction is toward reduction of articulations in lexical material. This reduction is characterizable in two complementary ways: as the *substantive reduction* of existing articulations, as in lenition processes, and as *temporal reduction*, in which existing muscular events progressively overlap in time, as in assimilation and related processes.

4. Notational and theoretical preliminaries

Some clarification of notational and theoretical devices is necessary at the outset, since our perspective on changes in articulation differs substantially from that assumed by structuralist and generativist models. Wherever possible, however, we will use the familiar devices of these models as a means of bridging the gap between theories, especially in

those areas in which overlapping interests but differing approaches demand careful attention.

4.1. Muscular events

We look to the muscular events comprising individual core lexical items as the sources of events which survive in subsequent manifestations of those items. That is, in our view internal change is lexically specific, unidirectionally reductive, and restricted by the timing and activity patterns of muscular events in individual lexical precursors.[8] We therefore view lexical items as complexes of muscular events which are temporally ordered, but not in the serial segmental fashion familiar from classical generative bundle of feature notation.[9] Vocal tract activity may be schematically represented as in Fig. 1.

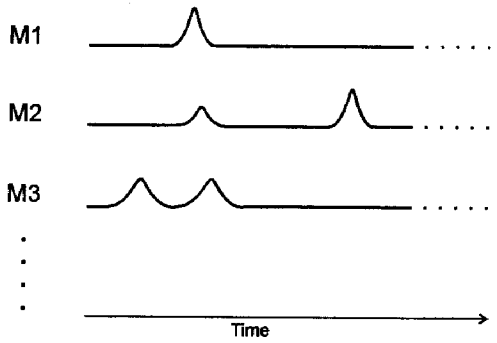

Figure 1.

In Fig. 1, articulation is expressed as if a multi-channel recording were being made of vocal tract muscular activity. In this schema, relational time is directly represented: one can see that one instance of the activation of M1 (i.e. a muscular event involving M1) partially overlaps (coarticulates with) M3, as does the first phase of activity of M2.[10] This schema also differs from feature representations in that rather than representing static or canonical vocal tract targets, it represents the muscular activity which precedes and produces vocal tract movement. Synchronically, muscular events exhibit constrained ranges of relative timing and magnitude within individual utterances, ranges limited by the extremes of the careful-to-casual speech continuum. This continuum is embedded in a larger range of highly

constrained unidirectional reductive changes in muscular event timing and magnitude – roughly, the diachronic phenomena constituting sound change.

Muscular events differ from the gestures of current articulatory approaches (e.g. Browman & Goldstein 1986, 1990, 1992) in the following way: gestures refer to control structures for the movement of articulators in spatial and temporal terms rather than to the bursts of muscular activity (muscular events) underlying any and all movements. Muscular events take place at particular times within utterances synchronically, and change in magnitude and timing diachronically. Spatial locations of articulators – the basis for both traditional static features and the more dynamic gestures of Browman & Goldstein – are derived, not primary, and, as we will argue, temporal and substantive reduction processes are not well-represented at the spatial level. Our conceptualization of reduction processes crucially depends on choosing a level of description at which reduction is potentially measurable across the entire range of articulatory activity. Common consonantal lenition patterns, which we discuss below, are *not* consistently characterizable as reductions in terms of either vocal tract shape or spatially-defined gestures, or on the acoustic level.[11]

In the coordinative structures approach, gestures refer to control structures for the movement of articulators, and the involvement of particular muscles will change with context. Thus the realization of, e.g. a [t] in a given phonetic context will involve whatever muscular contractions are necessary for that context (cf. Fowler et. al. 1980: 383): "[...] control over movement must be realized by bending the current field of forces in an intended direction [...]. An implication for speech production must be that there is no simple relationship between the distinctive features of a phonological segment and contractions of particular muscles".

If we abandon the assumption that standardized units function as building blocks and regard individual expressions as stored in a holistic fashion, then the contexts for all 'segments' internal to a given expression are constant, and firing patterns for particular muscles can indeed be specified. Also note that differential behaviors of the sort we identified in 2 are typically maintained under global increases or decreases in motor output, such as may be occasioned by variations in speaking volume, in the psychological state of the speaker, and so on. That is, even under varying conditions, the (a) and (b) versions of (1)-(4), for example, will still differ in the ways we have indicated. Thus, although various internal and external conditions may induce variation in motor output, expression-specific detail is preserved, not eliminated or normalized away. Such an approach also allows, for example, a given [t] (or its

gestural score equivalent) in a given context to change independently of what are assumed to be other, identical [t]'s in identical contexts. This is not to say that there are no units of the nature of [t]'s or [i]'s, but rather that they are not building blocks for core items, whose synchronic and diachronic behavior suggest that the locus of phonetic change is the finer level of expression-specific motor patterns. De-contextualized [t]'s and [i]'s, as well as larger structures, would be derived over holistic routines and available for other purposes, such as the production of non-core material, in which we expect standardization.[12]

Because schemata of the sort in Fig. 1 make accessible both substantive and temporal information – the magnitudes, relative onset, offset, and duration of muscular events – they afford more direct representation of assimilations and lenitions. For example, we might represent the regressive palatalization of apicals as a leftward displacement of the muscular events responsible for tongue doming, and the flapping of [t] as a diminution in the magnitude and duration of the muscular events involved in producing linguo-palatal contact.[13]

We note in passing that it was once hoped that a schema like that in Fig. 1 could eventually be derived from feature representations through the specification of low level phonetic output rules. That this derivation was never carried out is due not only to the difficulty of the undertaking but also to the prevalent attitude that nothing so low level could possibly shed any light on problems in either synchronic or diachronic description.

Though it rejects the rigid segmentalism of its parent theory, the autosegmental approach is limited by retaining the commitment to classificatory distinctive features which prevented its predecessors from directly modeling articulation. One reason this is not apparent is that an increase in realism is possible the more closely the classificatory feature required in an analysis corresponds to a single muscular event. Thus, when the phenomena of interest are describable in terms of e.g. [high tone] or [nasal], the classificatory nature of features causes little difficulty. Since very few of the features in general use approximate a single muscular event, however, the superficial advantage of multi-tiered representations – affording reference to articulations independent of segmental boundaries – is limited to those features which model phenomena such as nasality, tone, and lip rounding and protrusion which, though hardly attributable to single muscular events, seem to be relatively independent. But for phenomena involving lingual articulations, which result from the activity of muscular events which are just as independently controllable, albeit not so obviously, this system of feature representation provides no corresponding referential independence.

This discrepancy did not, however, foster a reconsideration of the utility of classificatory features for autosegmental analysis, and the adequacy of features like [coronal] and [anterior] for description on the place of articulation tier remains unquestioned (Lieber 1987). Thus, [coronal] remains, although several independently controllable muscles are involved in positioning the tongue in sounds labelled with this feature. The other place of articulation workhorse, [anterior], remains as well, even though it refers, not to articulator movement but to an arbitrary spatial division in the vocal tract.[14]

This is not to say that phonological investigators have neglected the muscular events which underlie speech. In an attempt to provide the long-promised low-level specification of distinctive features, Halle (1983) presents a model wherein traditional features are converted into muscular command specifications for segmental targets. Even suspending consideration of the question of whether, if such a specification were to succeed, it might not render the distinctive feature level of representation superfluous,[15] there remains a fundamental problem with this enterprise: bursts of muscular activity *move* the articulators along a *trajectory* and hence are more appropriate for describing the transitions between targets than the targets themselves. Enlisting the most dynamic aspects of articulation to describe the most static ends – targets – mistakenly equates movement with stasis.

Despite such attempts at remediation, the goal of a more faithful modeling of the phonetic basis of phonological phenomena will continue to be undermined by the primitives inherited from the parent theory. Features simply do not refer to, and so cannot be used to refer to, muscular events. By design classificatory in nature, they are rather referentially somewhere between gross articulator positions and the muscular events responsible for them. Equally inhibiting is the continued reliance on segments in all such approaches, since the relatively greater attention to detail these later versions of the theory have made possible is vitiated by the still requisite monitoring of alphabetic comings and goings. Diachronic accuracy is not served by representing any stage of the gradual erosion and absorption of a vowel, for example, as the instantaneous deletion of a V-slot (as in Lieber 1987:110).

We will argue that phonetic detail, and hence low level representations of the sort in Fig. 1, helps us to understand many diachronic processes, including some, like certain cases of epenthesis, which at first glance appear to have non-reductive aetiologies. Further, the representation of internal change in terms of a very restricted set of changes to patterns of muscular activity removes us from the uncomfortable position of trying to describe phonetic change as a sequence of failures to reach canonical segment targets.

Segmentation has imparted an inappropriately static character to synchronic and diachronic studies. For synchronic expression, the assumption of the existence of segments created the problem of reconciling them with the realities of continuous speech.[16] The parallel problem created for diachronic description is the necessity to postulate a period of stasis punctuated by abrupt segment mutation, when good evidence points to finely-graded change. In phonetic explanations of both synchronic and diachronic events, the problem of segmentation has spawned a pseudo-solution in the form of the target/transition dichotomy.

As difficult as targets are to justify synchronically, it is in the fluid domain of articulatory evolution where the notions targets, failure to reach targets, and the (constant) reanalysis and restructuring of targets seem most poorly motivated. These notions impose an arbitrary stroboscopic perspective on diachronic processes, somewhat akin to describing the long-term physical dynamics of a weather front solely in terms of its success and subsequent failure to be located over Cincinnati. Only in the most trivial sense can an articulatory change be regarded as a failure to remain the same.

In defense of target theory, a number of investigators have directly compared speech production to non-symbolic motor behavior. Attempts to describe articulation in terms of non-symbolic behavior or general theories of movement are of interest to linguists not only for their explanatory utility, but also because such unified theories are more powerful by virtue of their generality than domain-specific theories. Thus the phonetic realization of target points of articulation has been compared with the context-specific task of hitting a ball during a tennis volley (MacNeilage 1970). Proponents of coordinative structures theory compare the articulatory dynamics of speech to those of non-symbolic behaviors such as locomotion (e.g. Kelso et al. 1984) and present transient articulator loading evidence to show that articulatory targets in speech are achieved through compensatory articulation, just as in non-symbolic behavior. As attractive as these unified theories of action may be, there are reasons to believe that extending such comparisons and experimental results to normal articulation, and especially to the evolution of symbolic behavior, may not be valid.

First, from a synchronic perspective, the failure to reach targets (through 'laziness', undershoot or coarticulatory influence) is well tolerated and even expected in speech, but usually disastrous in tennis. Failure to reach a target in tennis occasions a complete realignment (or cessation) of activity which is absent in such failures during fluent speech. Similarly, the comparison of fluent speech to locomotion over an idiosyncratic terrain ignores the fact that the vocal tract is an integral and non-varying part of the speaker, not an external and unique

collection of obstacles. Finally, static or transient perturbations to articulation common to experimental protocols are not common in everyday speech, nor are nonsense syllables comparable to a language's set of overlearned core expressions.

From a diachronic perspective, symbolic behavior evolves at a much faster rate than non-symbolic behavior. One can hardly imagine the evolution of tennis into a game in which a missed ball would be casually ignored, but language change even over a couple of centuries typically exhibits many missed and ultimately lost targets. Evolutionary changes to animal locomotion also proceed very slowly relative to changes in symbolic behavior. Here again, reductive changes to locomotion synergies (e.g. partial or total loss of gait components) can have disastrous consequences which are not paralleled in the evolution of symbolic behavior. Experimentally induced disruptions of symbolic behavior, however valuable in elucidating compensatory mechanisms, have no direct bearing on the normal evolution of symbolic behavior. Since speakers do not generally encounter such disruptions, there is no guarantee that compensatory principles are generalizable to the diachronic domain.

The extension of disruption studies to diachrony is made even more difficult by the following consideration. One of the principal findings of these studies is that vocal tract targets are achieved even under conditions which should impair an articulator from reaching a target point of articulation. If these findings are extended to the diachronic domain, we would expect evolutionary changes to be very slow or even non-existent. That is, if compensation is so automatic and successful under relatively adverse conditions, it should be even more so under normal conditions, virtually eliminating any possibility of change. A quandary for theories of phonetic evolution grounded in the findings of these research paradigms is therefore to explain how any change in articulation occurs at all.

If the synchronic consequences and evolutionary time courses of symbolic behavior differ from those of non-symbolic behavior, there is reason to believe that these two types of action may not be governed by the same principles, and that attempts to unify them under a single theory of action will fail. If this is the case, then, just as we recognize a difference between the perception of speech and the perception of non-speech sounds, we may want to consider fully recognizing an analogous difference for production.

The rejection of target theory for synchronic analysis at once eliminates the artificial distinction between coarticulation which affects only so-called segment boundaries (the concept of segment boundary blurring; cf. Keating 1985a) and coarticulation which is conceived as causing a change in the target of an adjacent segment (assimilation).

Boundary blurring is regarded as a temporal effect while assimilation must be called a spatial effect because of target readjustment. But boundary blurring and assimilation differ only in the degree of temporal overlap of the muscular events comprising the articulatory stretches in question. As changes in the timing of muscular events proceed, they will occasion, epiphenomenally, a change in the trajectory and target of an articulator. When expression of articulation is by means of the activity and timing of the *effectors* of articulation rather than spatial configurations presumed to be phonologically relevant, the formally distinct phenomena of boundary blurring and assimilation can be unified.

The relevance of degree of temporal overlap to the main topic of this paper, reduction, is the following: certain degrees of temporal overlap, somewhere between those traditionally regarded as boundary effects and those viewed as assimilations, will sometimes be transcriptionally rendered as insertions of new segmental material, i.e. epentheses, where in fact only the temporal contiguity of old, already present, muscular events has changed. This will be discussed in greater detail in later sections; for now, we offer a schematic of the concept in Fig. 2.

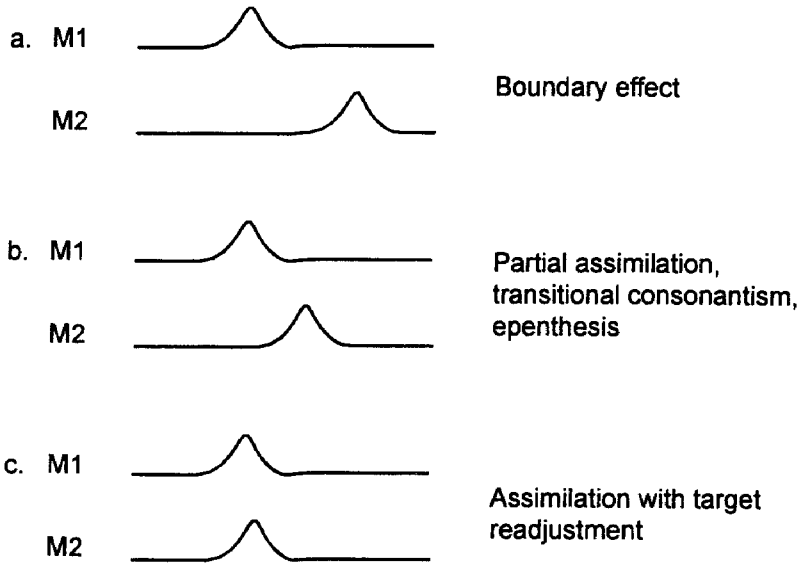

Figure 2.

In timing (a) the contiguity of the muscular events may be insufficient to produce anything more than a boundary effect, which would

probably not be represented even in narrow phonetic transcription. Timing (b) might well find expression in phonetic transcription, but the nature of its segmental representation might vary radically, all the way from low-level phonetic diacritic to epenthetic segment if the increasing contiguity of muscular events vector an articulator into near or complete contact with another articulator. Note that the variance in representation will be wholly an artifact of the segmental transcriptional tools at hand. Finally, timing (c) might be reflected by assigning a new label to one of the segments.

The schematic sequence in Fig. 2 can be used to represent arbitrarily-selected stages of a diachronic process of assimilation. Moreover, it could just as easily represent different degrees of assimilation in increasingly casual/allegro speech styles of a single speaker. The representation of casual/allegro speech as temporal and substantive reduction of existing muscular events is intuitively far more appealing in its simplicity than bringing the entire phonological feature-and-rule machinery to bear to transform old targets into new ones. This is particularly evident when one considers that feature-and-rule manipulation can theoretically bring virtually any target into prominence, being limited only by post-hoc naturalness and markedness constraints.

The claim that temporal and substantive reduction are sufficient to describe both internal change and casual speech phenomena and that no new articulations are introduced in such changes is also a hypothesis about the essential identity of these two classes of phenomena. In effect, it is a stronger version of familiar observations of the similarities between the processes of casual speech and commonly-encountered diachronic events, suggesting that in the changes evident in casual speech we have both a preview of some of the future of a given dialect and a window on the most frequent sorts of changes to phonetic form over time.

A similar approach is taken in the gesture-based phonology of Browman & Goldstein (1986, 1990, 1991, 1992). However, though their theory abandons the static segmental framework in favor of a more dynamic approach, it retains a commitment to target theory. Gestures are expressed as groups of vocal tract variables which are 'set' in order to achieve an articulatory 'task', 'goal', or 'target' (1990:343ff.). Although such gestures are expressed in physical terms (e.g. the vocal tract variables of tongue tip constriction location and degree), Browman and Goldstein's desire to relate casual speech tokens to canonical forms forces their notion of gesture onto a more abstract plane. While gestures are invariant on this abstract level, their execution in real time contexts occasions varying degrees of gestural overlap, which in turn are realized as 'context-varying articulatory trajectories' (1990:342). Though this is reminiscent of classical bi-level (phonological/phonetic) approaches to

description, two strongly stated claims show that the model has important implications:

> In faster, casual speech, we expect gestures to show decreased magnitudes (in both space and time) and to show increasing temporal overlap. We hypothesize that the types of casual speech alternations observed (segment insertions, deletions, assimilations, and weakenings) are consequences of these two kinds of variation in the gestural score. (1990:360)

> The gestural approach to casual speech is *highly constrained* in that casual speech processes may not introduce units (gestures), or alter gestures except by reducing their magnitude. (1990:371).

Very similar claims about the nature of the changes observable in both casual speech and diachrony are made in Pagliuca & Mowrey (1987:459, 466ff.) where gestures are defined as neuromuscular events. The major differences between Browman & Goldstein's proposal and the theory put forward here have to do with the composition of the independently controllable primitives which participate in reduction and temporal overlap and the range of phenomena the two proposals are intended to encompass. The substantive and temporal reduction hypothesis not only unifies the careful-to-casual continuum with more long-term diachronic events, but at the same time implicitly suggests extensive revisions of reconstruction methodology. Moreover, we expect the hypothesis to be valid for internal changes in non-speech modes of learned communicative behavior, including the signed languages of the deaf and facial and other kinesic gestures.

We have chosen muscular events over spatially-defined gestures as our independently controllable primitives not only because of the derived nature of the latter and the problems associated with target theory, but also because the longer-term continuity of reductive changes is describable in neuromuscular terms but not on the gestural level of description. With muscular events as primitives, the branching patterns so commonly seen in evolutionary change, which are not easily definable in spatial terms, are specifiable directly. We are also confident that consonant lenition patterns such as $t^h > t^s > s$ and $t^h > \underline{t}^h > \theta$ can be given neuromuscular specifications in accord with the hypothesis that these are indeed reductive events (see 5.3.2). In contrast, Browman & Goldstein's gestural model, as currently formulated, seems incapable in principle of providing a reduction explanation for these changes. For the change $t^h > t^s > s$ one might correctly say that the alveolar closing gesture {τ} weakens to become the alveolar near-closing gesture {σ} (1990:344). But this is a transformation of one gesture into another, not simply a

reduction in the magnitude of the gesture as the second claim states. More importantly, for $t^h > \underline{t}^h > \theta$, the vocal tract variable of tongue-tip constriction location progresses to a more extreme forward location with time. The {τ} gesture therefore either shows increasing magnitude (contra the first claim) or must be transformed into a new gesture (contra the second claim). Similar problems attend Browman & Goldstein's analyses of assimilatory phenomena, which we address in 5.1.

A theoretically more profound difficulty is encountered in extending Browman and Goldstein's proposals to the broader diachronic domain. If gestures are defined spatially, and may only reduce in magnitude and overlap in time, the long-term prognosis for phonetic evolution is what may be called the 'schwa catastrophe'. That is, if spatially-defined gestures can only diminish and overlap (which may occasion further reduction through hiding), eventually their diminution will lead to no spatial displacement of the articulators at all – the neutral vocal tract configuration (see 5.5).

Investigation of neuromuscular activity in speech is still in its infancy and therefore many of the claims we put forward here will rely on purely theoretical argumentation coupled with reasonable interpretations of the articulatory factors underlying the processes we will examine. Although some of these processes will be partially described using familiar segmental devices, our goal will be to show that these processes are best described as a highly restricted set of changes in neuromuscular activity patterns.

To summarize, our general hypothesis is that internal phonetic evolution consists of just two types of changes in muscular activity patterns: bursts of muscular activity become temporally more contiguous (the temporal distance between them is reduced) and diminish in amplitude (reduce in substance or energy content). Further, we propose that most currently recognized types of phonological processes can be reduced to one or both of these types (Pagliuca 1982:8,277; Pagliuca & Mowrey 1987). Note that in this conception neither segments nor features are relevant units. In effect, we are proposing a form-based or substantive rather than a segment-based or structural phonology, in which the focus is on changes to particular motor patterns in lexical and grammatical material, and not on the epiphenomenal changes to classificatory inventories and rule systems which sometimes appear to result from them.[17]

4.2. Internal and external change

In our view, a realistic theory of internal or evolutionary change takes as its starting point the evidence for the overwhelming predomi-

nance of reductive changes, in effect treating it as the primary datum to be accounted for. The theory should offer a way of analyzing the kinds of changes identified as reductive in as revealing and comprehensive a fashion as possible, such that we can ultimately address the question of why change should be reductive, rather than augmentative or directionless. But it must also provide some account of apparently intractable data – those changes which appear not to be reductive, and which have traditionally led analysts to dismiss out of hand the plausibility of a general theory of change of the sort we are proposing.

Among readily-identifiable types of ostensibly recalcitrant data are: 1) changes which result in the apparent insertion of entire segments, such as vowel epenthesis and prothesis and glide, fricative, and stop intrusion; 2) fortitions, including consonant gemination and manner changes which result in increased closure (and, by some accounts, affrication and other changes which 'add' stridency); and 3) apparently unconditioned changes of consonantal place of articulation. Even so apparently trivial a change as a labiodental fricative becoming a bilabial is augmentative, as it involves the addition of labial muscular events. Such changes are generally regarded as problematic for a view of change as articulatorily reductive, as all appear to be augmentative, adding either an entire segmental stretch, the supplemental energy necessary for increased constriction or increased closure duration, or an acoustically prominent property. To these types, which have traditionally been cited as prima facie examples of the necessity for assuming that change cannot be regarded as solely reductive, may be added metathesis, which is also often pointed to as evidence disconfirming a view of change as exclusively gradual and reductive.

In section 1 we indicated that at least some apparently unconditioned place of articulation changes which have been regarded as necessarily perceptually-induced appear rather to be the result of temporal and substantive reduction, and do not involve the replacement of a complex of original muscular events by new events. Similarly, affrication, defined as a strengthening by Bailey (1977), appears instead to represent the decay of stop closure integrity, which, if it continues, leaves simple fricatives. $\phi > f$, also sometimes regarded as a strengthening because of the greater acoustic energy associated with stridents, rather seems to be principally the result of the reduction of activity of the orbicularis oris superior; here, as in affrication, a decrease in articulatory energy has as a byproduct an increase in acoustic energy.

Also potentially describable as reductive are certain kinds of consonant gemination, especially those in which geminates arise as a consequence of the absorption of remnants of contiguous segments, and, as suggested above, certain kinds of metathesis. Vowel epenthesis and

intrusive consonantism, which we address in this paper, as well as manner changes resulting in increased closure, which at first glance appear to be more problematic, are also in principle amenable to reduction analyses.

In the evaluation of apparent anomalies, we first want to know whether the change at issue is attested or reconstructed. In the latter case, the evidence is not of the most compelling sort, and if an alternative plausible reconstruction consistent with our principles can be provided, there is insufficient reason to abandon the hypothesis. Secondly, we would like to be reasonably certain that the change is internal; if it is the result of language contact, then it is not within the scope of what our theory proposes to explain. In such cases, it is sufficient to show that there is reason to believe that the change was not internal. For the remainder, i.e. for attested internal changes, the approach must be to attempt to show, for at least some of the kinds of phenomena represented, that their anomalous status is only apparent, and that they too can be regarded as reductive changes. For every change type removed from the list of diachronic events which appear to be non-reductive, the proportion of such anomalies is reduced and the plausibility of the general hypothesis thereby increased.

We will try to show that the theory of change outlined here offers a uniform and comprehensive view of phenomena generally recognized to be reductive. Given the tendency to reject reduction-only views of change, however, we will also try to demonstrate its potential in handling more difficult cases. In particular, we will examine representative instances of one class of events which ostensibly invalidate our general hypothesis by appearing to run counter to the statistically predominant tendency toward reduction. If compelling evidence that such events are frequent in evolutionary change should be uncovered, then our assertion about the nature of change and the benefits of pursuing it could not stand unmodified. We contend, however, that for most of the changes at issue what we are seeing in those cases which are not externally-induced is the substantive and temporal reduction of existing articulations.

The events in question are epentheses and protheses. For expository purposes, we broadly define these phenomena as those in which an entire segment appears to have arisen where there formerly was none, and for convenience call all such events epentheses. Broadly speaking, two classes of cases are recognizable. In one class, in changes of the type *AC > ABC, B,* which appears to be an epenthetic element, is in fact not intrusive articulatory material, but rather consists of muscular events already present in the prior stage *AC*, albeit now segmentalized differently. In the second type, the evidence for the temporal priority of *AC*

over *ABC* is inferential rather than direct, as, for instance, when *AC* is reconstructed rather than attested as the ancestor of *ABC*, or when an attested *AC*'s status as the lineal parent of *ABC* is assumed rather than proven. In such cases, we argue, it is possible that what we are seeing is differential retention of *B* across dialects, with *ABC* > *AC* the actual historical development. We treat phenomena of the first type in section 6 and phenomena of the second type, which are much more pervasive in historical research, in 7. To make our overview of epenthesis appear more sensible to the reader, we begin in 5 with a discussion of temporal and substantive reduction.

We will here restrict our discussion to *internal* or *evolutionary* change. The term *internal* will roughly correspond to those changes which are thought to be phonetically motivated within a dialect. Excluded are changes which are obviously induced by contact between distinct dialects or languages, as well as hypercorrection and abrupt changes to archaic or infrequent forms (Weinreich 1953:57; Malmberg 1963:101f.; Anttila 1972:187). These latter all behave in much the same way as lexical material borrowed from a different language, and we group them together under the terms *external* or *non-evolutionary* change. In doing so, we follow in the tradition of theories of change which are careful to separate analogical changes and borrowings from processes which do not seem to require external motivation. Naturally, the potential for abuse of this dichotomy is great, since any change which does not fit a theory can be neatly disposed of by casting it into the trash heap of analogy and contact. The theory we present here cannot entirely avoid this danger, and there will be cases which cannot be decisively resolved.[18] Nevertheless, clear cases of borrowed vocabulary and hypercorrection show the usefulness of a division into internal and external change, especially along the dimensions of phonetic explicability and rate of change.

Thus, for example, non-evolutionary epenthesis may arise when the articulatory fit of alien consonant clusters onto the borrowing language is so poor that the best approximation is a pattern containing vocalic articulations absent in the foreign model. Just what this articulatory pattern will be depends on pre-existing patterns, which vary from one borrowing language to another. Consider, for example, the differential treatment of English *stay* uttered by a Vietnamese speaker and a Spanish speaker. The Vietnamese speaker will typically resyllabify [stej] as [sit te], with a short vowel with high rising tone between the initial [s] and [t], whereas the Spanish speaker will render it as [estéj], with a prothetic vowel. Neither Vietnamese nor Spanish has [st]-like patterns in word-initial position. For Vietnamese, the best approximation is [sit], since the short high front vowel is the closest transitional

articulation between [s] and [t]. This vowel takes a high rising tone because vowels preceding non-glottalized voiceless consonants are always high rising in tone. Spanish, on the other hand, has a pre-existing articulatory pattern containing medial [st] with a short unstressed vowel preceding the [s]. In neither language are new muscular event patterns introduced when *stay* is borrowed. The new word is made to conform to old articulatory patterns – no new epenthetic patterns have developed in the borrowing languages. Accommodation of this sort is not restricted to language and dialect contact in the usual sense: even within a given language, consonant clusters which are rare or completely absent in native core vocabulary may be subject to epenthesis by accommodation to engrained motor patterns.[19]

Accommodation in borrowing or language adoption can also result in fortitions and apparently unmotivated changes of place of articulation. Thus, fricatives may be rendered as aspirated stops if the latter but not the former occur in the adopting language. For an [f] in a target language, a [p^h] may be substituted if [p^h] but not [f] occurs in the adopting language. Just as non-remarkably, an [f] may be substituted for [p^h] if the situation is reversed, but only this, and not f > p^h, would be expected in internal evolution. Glides, too, may 'become' stops, as in the familiar case of the importation of [w]-initial Frankish words into Vulgar Latin dialects of France and Italy (OFr *gwarder*, *guerre*; It. *guardare*, *guerra* (Pope 1934:227f.; Grandgent 1927:67f.)). Even native words, if they are rare or otherwise peripheralized, may be subject to accommodation, as in the 'change' of syllable-final [p] to [k] in colloquial varieties of Spanish. The rate of occurrence of syllable-final [p] in Spanish is vanishingly small (Navarro 1968:17); virtually all occurrences are in learned words ('cultismos') which (re-)entered the language as borrowings from Latin long after original, inherited syllable-final [p]'s had decayed to labial glides or entirely, as in *cautivar* 'to captivate, charm' and *escrito* (< Lat. *scrīptu*) 'written'. As it is absent from the core lexicon of Spanish, it is likely that many speakers never encounter, and hence never lay down motor patterns for, syllable-final [p]. Consequently, cultismos such as *adaptar*, *captar*, *concepto*, etc. are in effect not native, but as alien as the English borrowing *Pepsi*, and all such items are approximated with the more frequent [k], which does occur syllable-finally in core lexical material, in place of [p]. Note that such changes are necessarily phonetically abrupt. An analogous case in Macedonian, resulting in the apparent change x > k, is discussed in Pagliuca (1982:224, 264f.), as is the change of Latin velars to Rumanian labials (Lat. *noctem*, *cognatu*: Rum. *noapte*, *cumnat*), which is argued to be an example of accommodation in language adoption (pp. 217-22).

Mindful of its incompleteness, we summarize our conception of internal vs. external change in Table 1.

Table 1.

	Internal change	*External change*
Domain:	Core vocabulary	Peripheral vocabulary, 'cultismos', borrowed vocabulary
	Frequent lexical material	Infrequent lexical material
Origin:	Physical	Extraphysical/psychological (analogically motivated)
Most easily observed in:	Relatively unmonitored speech	Monitored speech, hypercorrect speech, reading pronunciations
Time course:	Gradual	Abrupt

5. *Temporal and substantive reduction*

5.1. *Two views of assimilation*

We begin by considering the nature of assimilation, one of the most commonly-encountered types of diachronic change. Suppose the event of interest is an instance of regressive assimilation, and let it be represented schematically as $XY > X^{y}Y$, where X and Y are segments, and X^{y} indicates that X has taken on some property or properties of Y (is now a Y'ed or Y-influenced X).

There appear to be two typical ways of interpreting the event which intervened between the stages XY and $X^{y}Y$. The first may be called the phonological view, according to which X, which may be a classificatory element of either the phone or phoneme sort, becomes more like the element Y by simple contiguity to it. On this view, the y-ness of X^{y} is not originally part of Y, i.e. the physical source of ^{y} of X is not Y. Though Y may be regarded as having instigated the rise of y-like properties in X, it is not viewed as having lent, as it were, some of itself to X. This is not to say that the properties in ^{y} of X are not physically like the relevant

properties of Y, but only that their provenience is not in Y. In fact, on this view, their provenience is immaterial and not at issue.

The second view may be called the (articulatory) phonetic view, in which the change is regarded as having involved some of the articulatory events associated with Y 'spilling over' into the temporal domain of X, affecting X such that its articulation now anticipatorily includes part of the articulation originally confined to Y. On this view, Y *is* the material source of the articulatory event in ^{y} of X^{y}.

Regardless of which view is adopted, assimilation phenomena are represented by phonologists with distinctive features, one of the principal advantages of which is supposed to be their utility in portraying phonological processes. We may therefore ask to what extent such representations accurately model assimilatory phenomena. We begin by asking whether features, which are primarily classificatory devices, can do a serviceable job as stand-ins for the articulatory properties relevant to assimilatory changes. Let the lower case letters in (8) stand for articulatory properties and ***b*** that property which characterizes the y-ness of Y that becomes associated with X to result in X^{y}.

(8)

(X)	(Y)		(X^{y})	(Y)
a	a		a	a
	b	>	***b***	***b***
c			c	
	d			d

Faithful to the traditional interpretations, (8) portrays the assimilation event as consisting solely of the copying or spreading of b of Y onto X. That is, it suggests that the post-assimilation Y (the Y of $X^{y}Y$) is its old self (the pre-event Y of XY) and X^{y} is its former self (the X of XY) *plus* b. But this greatly reduces the plausibility of interpreting properties a thru d as actual articulatory events, or as anything other than static classificatory elements. In particular, a thru d cannot refer to muscular events, which have magnitudes and time courses: for (a given magnitude of) b to remain in Y and also appear in X would require a doubling of the energy associated with the original event b – an event of the same magnitude at roughly twice the temporal duration. This would mean that assimilation is in essence augmentation, with muscular events accruing to X without reducing in Y, with the result that the articulatory density of the original sequence XY is increased. Because over time subsequent assimilatory changes may be expected to occur in the same lexemes, to interpret assimilation in such a manner is to suggest that it always moves linguistic material in the direction of infinitely dense motor activity. Since this is not what happens – since, on the contrary,

over time the individual lexical items that assimilations affect typically display eventual reduction in length and articulatory density – something crucial is missing from the representation of assimilation as copying or spreading, which forces assimilatory changes to be portrayed as augmentation.

As a first approximation, a more plausible interpretation of assimilatory phenomena is that they involve the migration of the *original muscular event b* of *Y* into the temporal domain of *X*. That is, rather than a spreading or local increase in energy and hence overall density, assimilation involves a change in the temporal location of the particular energy in *b* relative to the other muscular events in the original stretch *XY*. This would mean that neither *X nor Y* in the 'after' stage is identical to its antecedent in the 'before' stage, and that, generally, in assimilatory events *X* and *Y* are jointly affected.

The difficulty with the first approximation is that temporal migration – temporal reduction – alone seems insufficient to account for observation. If assimilations consisted solely of the migration of existing muscular events, then we should see maintenance of existing energy for the 'before' and 'after' stages of individual items in which assimilatory events had taken place. That maintenance is unlikely is perhaps not obvious if we think of single historical events. But when we consider that over time – after multiple regressive assimilatory events have occurred – it would entail a local density increase in the form of a piling-up of this energy at the leftmost boundary of individual lexical items, the limits of a 'temporal reduction only' view of assimilation become apparent. The characterization of assimilation as temporal reduction must, then, be fundamentally incomplete. Assimilation must involve or be accompanied by some kind of substantive reduction.

Precisely where and how the substantive reduction takes place – whether it involves muscular events upstream or downstream of the migrating event(s), is simultaneous with the temporal migration or not, and so on – are questions which require answers which can only come from properly designed instrumental study. Nevertheless, some indications of what is likely are gleanable from a consideration of the typical life histories of the more common sorts of regressive assimilation.

Regressive vowel nasalization, for example, is often the precursor to an eventual loss of the nasal consonant which conditioned the nasalization. This suggests not only that the muscular events which decay (undergo substantive reduction) are the lingual and other non-nasal components of the original nasal consonant, but that they do so at roughly the same time that the velopharyngeal events migrate. Similarly, in phenomena involving lingual muscular events, such as palatalization, at least some of the events which decay are likely those originally

associated with the conditioning vowel which do not participate in the coproduction which palatalizes the consonant. Distant assimilation should be different only in that the migrating events have farther to travel and would necessarily affect the intervening material. Thus, the fronting and/or raising of vowels in some kinds of umlaut phenomena is also likely brought about by the temporal migration of events originally part of the conditioning vowel, leaving behind remnants which then erode.

That there exists an intimate relation between assimilation and loss has not gone unnoticed by diachronists. Ohala, for example, has observed that "[...] it is often the case that sound changes due to assimilatory processes, as in [...] a*n* > ã; fo*ti* > føt involve the apparent simultaneous loss of the conditioning environment". (Ohala 1983:256f.). For Ohala, however, assimilation or overlap of articulations merely sets up the conditions for a speaker's reinterpretation or reassignment of properties.

If we are correct in asserting that the temporal reduction which gives rise to assimilation must be accompanied by substantive reduction, then observed associations between assimilation and reduction and loss are precisely what we should expect. But in conventional notation, the relationship between assimilation and reduction and loss is anything but obvious. One reason for this is that mutations will typically be recorded only when a change has proceeded far enough to warrant a change in the available inventory of diacritics or segments. Even then, our notational systems force us to represent the *Y* in $X^{y}Y$ as its old, still-full self, recording only the modification of *X* to X^{y}. They also force us to speak in terms of 'conditioning' vowels or consonants, as if their contribution were instigatory, whereas, in e.g. *foti* > *føt*, it is not that the relevant articulations of the final vowel *condition*, but rather, by migrating, materially *constitute* part of the change. In the traditional cause-and-effect depiction (but not in Ohala's view, nor that of Browman & Goldstein, to which we return below), the umlauting occurs entirely independently of the subsequent loss of the conditioning vowel, which is sometimes characterized as being deleted because it had become redundant. The usual representation of the intermediate stage (following the fronting/raising, but prior to the deletion of the conditioning vowel) therefore cannot help but suggest that the conditioning vowel has survived the umlauting intact. We suggest that the vowel which deletes must, because part of its substance has gone to raise the vowel to the left, be but a remnant of its former self. Similarly, in vowel nasalization, only when concomitant substantive reduction has proceeded far enough are the necessary notational adjustments made – deletion of (the remnants of) the original nasal consonant responsible for vowel nasalization.

Similar notions of conditioning followed by deletion of the conditioning environment are found in Browman & Goldstein's (1991) account of the Middle English change of [x] to [f] in words such as *cough* and *tough*, but here the process appears as 'reassignment' followed by subsequent deletion (p. 328), and is treated as a perceptually-mediated change given certain articulatory preconditions.

Because no basis for articulatory conditioning is evident, this change has been argued to require recourse to acoustic or auditory factors; Ladefoged (1971) and Hyman (1973) interpret it in terms of the substitution of one [grave] fricative for another, and hence as evidence for the importance of acoustic factors in synchronic and diachronic analysis. An analysis of this and similar changes as articulatory in nature is outlined in Pagliuca (1982:169ff.); a similar analysis for the Middle English change is proposed by Catford (1991:177).

Certain details on this change are available; the evidence in Dobson (1968), a study based on the records of the English orthoepists of the time, is particularly instructive. Dobson's account makes clear that all items undergoing this change consisted of a diphthong with a labial offset followed by [x], and, moreover, that the diphthong was reduced if the change to [f] occurred, but otherwise remained intact (in dialects in which the change did not take place, and in words in which [x] decayed to [h] and then [∅] (p. 506 et passim)). Crucially, the loss of the labial offset of the vowel (leaving a monophthong) was clearly contemporaneous with, and not subsequent to, the appearance of the labial fricative. Note that the contrast in the fate of the diphthong according to whether the change occurred or not by itself suggests that a misperception analysis cannot be correct: substitution of one [grave] fricative for another should have produced e.g. [ouf] from [oux] directly, without the observed concomitant reduction of the diphthong.

Citing Dobson's evidence as presented in Pagliuca (1982), Browman & Goldstein treat orthographic <*ou*> in these items as representing diphthongal articulations with gestural narrowing occurring over the latter part of the diphthong for both the velic and the labial constriction locations. They argue that the constriction degree parameter value associated with frication for the following [x] is reassigned to the diphthongal offset, "presumably on the part of the listener" (p. 328), because of increasing overlap of the velar fricative with the end of the diphthong. The velar fricative is subsequently deleted, leaving [ɸ] following the now-monophthongal vowel nucleus.

In our view no such listener reassignment is necessary if one additional property of fricative-like articulations is taken into account, namely, glottal opening (a muscular event involving the posterior cricoarytenoid). We agree on the characterization of the diphthongal

offset as 'labial narrow', but regard frication of the labial narrow vocalic offset to be a direct result of migration of a surviving glottal event associated with the decaying [x] articulations, i.e. as a laryngeal assimilation process with concomitant reduction of supralaryngeal muscular events associated with [x]. The increasing airstream velocity during the labial narrow phase would then be sufficient to produce the turbulence necessary for [ɸ].

The modeling of this process as a perceptually-based reassignment of constriction location demonstrates the inherent unsuitability of spatial target-based models for the description of certain assimilatory processes, especially those which superficially appear to create new targets. Crucially missing from the gestural framework as currently formulated are adequate principles of interaction for tract variables in adjacent gestural structures which are undergoing reductive processes. However, insofar as the gestural model's independently controllable motions through space do not reflect the independently controllable muscular events which underlie all motion, spatially-defined reduction processes will perforce appear to be unprincipled and in need of perceptual/psychological remediation.

5.2. Temporal reduction

If coarticulation changes come about by the temporal compression of muscular events, then the event-type of interest in assimilation is temporal reduction – or, more accurately, the interaction between temporal and substantive reduction. To illustrate, let us consider two relatively common diachronic assimilatory processes: regressive vowel nasalization and regressive palatalization of apicals.

(9) ãN > ã^{N}

(10) si > s^{j}i > ʃi

In (9), increasing nasalization of the vowel over time[20] takes place by a gradual retiming of a subset of the nasal consonant articulations – those muscular events responsible for velum lowering (palatoglossus activity and reciprocal levator palatine inhibition). Similarly, in (10) palatalization involves retiming of tongue-doming events, involving genioglossus and perhaps intrinsic lingual muscles. Thus, in each case, what is occurring is the temporal migration (temporal reduction) of a few muscular events onto an earlier portion of the articulatory sequence. Again, we suspect that concomitant with this migration is the substantive reduction of the non-migrating muscular events associated with the nasal consonant or high vowel. This would account for the observed

tendency of vowel nasalization and regressive palatalization to eventually entail the apparent complete erosion of the nasal consonant or absorption of the conditioning palatal vowel, respectively, both of which are familiar from the history of Romance and other Indo-European subfamilies.[21] Obviously, the gradual nature of these substantive and temporal reductions cannot be adequately captured using familiar notational devices, which portray phonetic change as lurching from one sequence of phonetic symbols to another. Thus the failure of substantive reduction to be recorded as concomitant with temporal reduction is in part due to the tendency for such reduction to be noted in transcription or orthography only when it is 'complete'.

Assimilations with concomitant substantive reduction are not limited to the supraglottal vocal tract. In much the same way as final nasal consonant reduction and the nasalization of a preceding vowel can be seen to proceed in tandem, the reduction and eventual loss of final obstruents and the genesis of tone seem to be very closely linked in many languages of Southeast Asia.

It has long been recognized that in many tone languages, consonantal voicing qualities pattern with tones on adjacent vowels. In addition, some of these languages, especially those of Southeast Asia, exhibit strong restrictions on the number and voicing qualities of consonants which can precede or follow tone-bearing vowels (Maran 1971a, b).

The linkage of voicing and tonality is intuitively appealing since both phenomena rely at least in part on laryngeal muscular events. But the simple binary feature [voice] is inadequate to describe the range of tonal phenomena, and rather than accept the revision of the feature system as proposed in Halle & Stevens (1971) (see also Halle & Clements 1983), most researchers in the field have been content to add new features specific to tone. This was an unfortunate choice, since the interaction of tone and consonant voicing quality could now only be described in terms of conditioned feature-changing rules rather than as assimilatory phenomena. We suggest that Halle & Stevens were on the right track with their proposed revision of the feature system and further that a description in terms of the muscular events responsible for laryngeal adjustments during consonant production will reveal tonogenesis to be an assimilatory process with concomitant substantive reduction of supraglottal articulations.

Since Southeast Asian languages show a patterning of consonant voicing and tone for both initial and final consonants, a controversy has arisen between those who ascribe tonogenesis to initial consonants and those who adhere to the position that loss of finals is the primary causal factor. Under the theory we present here, the linkage of final consonant reduction and tone suggests that it is the final consonants which impart

a residue of laryngeal muscular events onto the preceding vowel. If this is the case, we would expect this residue to perturb the offset of the vowel more than the beginning of the vowel, at least in the earlier stages of tonogenesis. Vowels acquiring laryngeal residues from final consonants would therefore have tonal contours which show the greatest perturbations on the final part of the vowel. Tones arising from *initial* consonant laryngeal residues should show exactly the reverse pattern. We can therefore appeal to instrumental studies of tone to decide the issue of tonogenesis in many cases.

One such instrumental study is Han (1969), wherein Vietnamese tones produced by speakers of Hanoi dialect are examined. The Hanoi dialect has six tones. Of these six, one is level and in approximately the high mid range, i.e. it is essentially non-tonal. For the five remaining tones, all show the greatest tonal excursions during the mid- to end-points of the vowel. Tonal differences at the onset of vowels are much smaller across the four informants Han studied. This phonetic evidence alone supports a post-vocalic consonant origin for Vietnamese tones over the pre-vocalic alternative. Restrictions on final consonants point to the same solution: only nasal and stop consonants are permitted in final position. If the final consonant is a stop, the preceding vowel must have one of two tones – the rising tone or the 'drop' tone (glottalization at the end of the vowel).

There is of course no reason why remnant laryngeal events should not encroach on the vowel *onset* given sufficient time. At this stage, oral muscular events associated with final consonants would probably have reduced even further, perhaps to the point at which no final consonants are evident. If the laryngeal remnants we call tone move sufficiently close to the vowel onset the voicing quality of pre-vocalic consonants should begin to change.

Consonantal voicing, as it is currently conceived, is, after all, primarily a measure of consonant release to vocalic voice onset time. If tone began to 'condition' initial consonant voicing, it would be difficult if not impossible through synchronic distributions alone to distinguish this situation from one in which *initial* consonants condition tone. This may be one source of the initial vs. final consonant tonogenesis controversy (Maddieson 1974; Hyman & Schuh 1974; Hyman 1976).[22] For Southeast Asian tone languages, in which final consonant reduction is so prevalent, the hypothesis of concomitant temporal and substantive reduction points to final consonants as the origin of tone.

The parallel between the processes of tonal development, palatalization and vowel nasalization is at least suggestive of a unified approach to all such phenomena, involving substantive reduction of the majority of the articulations originally present and concomitant temporal reduc-

tion of remnant articulations. In such a conception of phonetic change, which seeks to provide the basis of a uniform account of all temporal and substantive reductions, remnant articulations may come from any independently controllable neuromuscular entity in the articulatory apparatus (Pagliuca & Mowrey 1987:467). Exactly what is independently controllable is an as-yet unresolved empirical matter. Also, as the discussion around Fig. 2 above indicates, one of the advantages of the view we are advocating is that it allows us to bridge the descriptive and theoretical gap between synchrony and diachrony: viewing coarticulatory phenomena of both the (static) overlap type and the (dynamic) assimilatory sort as merely different in degree rather than kind allows a uniform treatment of synchronic and diachronic phenomena.

5.3. Substantive reduction

5.3.1. Vowel reduction

Now consider the nature of substantive reduction. Let us take as an illustrative case the reduction of a weak, unstressed interconsonantal vowel to zero, and represent it schematically in the usual way as *...CVCV... > ...CCV...*. In the traditional view, this event involves the elimination of a segmental unit from the phonological or phonetic program. Although the segmental and suprasegmental environment would be regarded as conditioning the deletion event, the segments which survive are not regarded as otherwise directly involved. In particular, the consonants which now abut do so merely to close the gap, as it were, left by the exiting of the vowel – they are not, for example, viewed as being now contiguous as the result of having gradually encroached upon the temporal domain of the weak vowel, ultimately engulfing and absorbing its remnants, though we suspect that this is likely much closer to the truth.

Once again, we can frame the problem in terms of temporal relations. Assuming that there is no need to consider the possibility that the post-change stage is of greater temporal duration than the original sequence, we have the following two possibilities: either the temporal duration of the new stretch *CCV* is equal to that of the parent sequence *CVCV* or it is less than it. But if the temporal durations of the two sequences are equal, then either there is a gap in the articulation of the new sequence (a period of silence where the vowel formerly was) or the surrounding consonantal material has expanded to encroach on the former vowel's temporal domain. Of these two, the only reasonable possibility is the latter, but the nature of the expansion could itself be either temporal or substantive: one or both consonants could display either the same muscular energy they originally had spread out over a

longer time, or an increase in energy, with one or both expanding at that edge which originally abutted the vowel, to fill the gap.

Again, given the evidence of the historical record, the latter possibility – energy increase – seems unlikely, as it would predict, if not augmentation, then stasis over time for the articulatory energy in particular material. It is the former possibility – the diffusion of existing articulatory energy attendant on, or as a concomitant to, the loss of a weak vowel – which must be evaluated relative to its rival, namely, that in phenomena involving substantive reduction (the loss of muscular energy) there occurs a concomitant diminution in temporal duration. But the history of individual lexical items in well-attested languages and language families (compare the forms of the cardinal numbers in Romance languages with their Latin ancestors, or in modern Germanic languages with Gothic, Old English, Old High German, Old Norse and Old Frisian, or any set of descendant lexical items with their antecedents) seems to rule out the diffusion alternative, as does a consideration of what must be true of articulation over time independent of the history of particular lexical material, which we treat briefly in 5.5.

5.3.2. Consonantal decay

Like vowel reduction versus epenthesis, instances of consonantal decay greatly outnumber those of consonantal strengthening and excrescent consonantism. Examples are so abundant in the literature for both remote and ongoing changes that we will not attempt to document this tendency here, but rather confine ourselves to an excursion into the mechanism underlying consonantal decay phenomena and a discussion of the theoretical importance of the imbalance.

Consonantal decay is usually discussed in terms of an overall measure of segmental strength, the strength/sonority hierarchy (Sievers 1901, Hooper 1976a), which has found a place in divergent views of change ranging from the abstract and phonetically minimalist Foley (1970) to the concrete and phonetically based Hooper. We will here recast the strength/sonority hierarchy in terms which are tied closely to observable phonetic properties but which conform to our non-segmental framework.

Although we have used the terms consonant and vowel and the symbols C and V freely for expository purposes, it should be obvious that a rejection of the segmental framework entails a rejection of any subcategorization of segments. In its place, we refer to stretches of articulatory behavior roughly associated with traditional consonants as *local articulatory maxima* in order to indicate that a relatively large number of muscular events have come into temporal coincidence through the processes of temporal and substantive reduction. At the other

extreme, reduced vowels and highly reduced consonants can be viewed as temporal stretches during which relatively few muscular events take place.

This viewpoint differs from conventional wisdom only in that it is stated in non-segmental terms: the relationship between segment types and the strength or energy of their underlying articulations has been well explored in terms of both speed of articulatory movement and the amplitude of electromyographic (EMG) signals from the musculature which produces such movements. Harris (1974:2294) summarizes these findings as follows: "From what we know about the way which the EMG signal is related to the resulting articulatory motion, a larger amplitude of EMG signal is translated into a faster acoustic transition and/or a more extreme articulatory position". Relatively extreme and rapid movements of the articulators are more characteristic of consonants than vowels.

The mapping of local articulatory maxima onto the strength/sonority hierarchy is as in Table 2.

Table 2.

Strength / sonority	*Traditional segment type*	*Local articulatory maxima*
Strong or non-sonorant	Voiceless stops	Temporal coincidence of many muscular events and/or high amplitude events
	Voiced stops / Voiceless fricatives	
	Voiced fricatives	
	Nasals	
Weak or sonorant	Liquids	Temporal coincidence of few muscular events and/or low amplitude events
	Glides	

This mapping is intended only as a rough guide; we do not yet have any detailed view of the muscular events which typically correspond to vowel-like versus consonant-like articulations. Consonants and non-reduced vowels may differ primarily in the way articulatory effort is distributed rather than in the amount of overall energy which each involves. Another problem with any such categorization is its implied universality. As Ladefoged (1980) has pointed out, consonants transcribed with a given symbol may differ significantly in phonetic properties from one language to another. Even casual observation suggests that total energy may be one dimension along which we find cross-linguistic variation; compare, e.g. the difference in 'strength' between the voiceless velar fricatives of Dutch and Spanish, both of which are transcribed with [x]. Even within a given language, one symbol may be employed to transcribe articulations which may differ substantially in strength. For example, the [p^h] in *potato* ([p^hət^{h}éjDo] or [p^hət^{h}éjDə]) is not as strong as the [p^h] in *putt* ([p^hʌt]); low stress in the first syllable of *potato* is realized not only as reduction in the vowel but also as weakness in the labial and laryngeal muscular events of the consonant. Dialectal and stylistic variants often display further laryngeal reduction, but the gradual erosion of labial closure is not well represented, appearing only as an abrupt deletion in the most reduced form; cf. (11). As we argued in 2, differential behavior is evident even when stress patterns and segmental environment are identical. For example, the first syllable of *particular* is typically temporally and substantively reduced relative to that of the ostensibly identical syllable in the less-often uttered *particulate*.

(11) [p^hət^{h}éjDə] > [bət^{h}éjDə] > [pt^{h}éjDə] > [t^{h}éjDə]

Fig. 3 shows typical decay pathways of voiceless aspirated stops expressed as segmental mutations. Alternative decay pathways are represented as branches from a higher node on the vertical axis. The parallel between consonantal decay pathways and the strength/sonority hierarchy is immediately obvious: in general terms, liquids, glides, and 'secondary' vowel or consonant qualities (represented in Fig. 3 as Q) descend from weakly-stopped or fricative configurations, which in turn descend from more fully stopped configurations.[23]

Given such common decay pathways, can we express the intuitive notion of strength reduction (most obvious at the extremes of the cline) in potentially measurable terms? Certainly the generative feature device is no help here: all surface segments receive a full binary specification right up to the point where C > ∅. For intermediate stages, no loss of strength can be discerned. Nor do acoustic measurements help. On the contrary, acoustic energy often increases as articulatory strength decreases. Loss of *articulatory* content, on the other hand, is obvious if we simply look at what is happening to the speech musculature as we move down the strength scale. For the labial decay pathways, this requires no sophisticated instrumentation: production of [p^h] requires muscular activity associated with both the upper and lower lips together with glottal opening. Production of [f] requires diminished or absent muscular activity associated with the upper lip while retaining lower lip

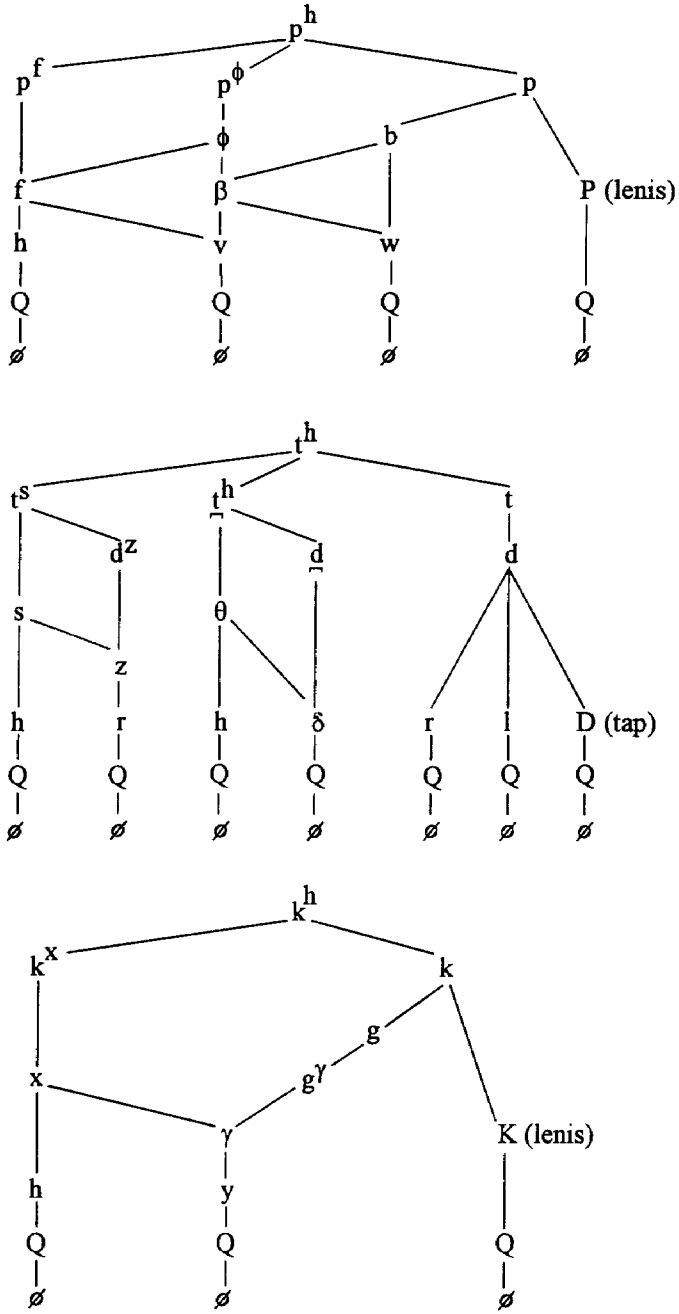

Figure 3.

and glottal opening activity. [h] requires absence of labial muscular activity but retains glottal opening activity. Finally, ∅ is the loss of all muscular events originally associated with the consonant.

Were such coarse muscular events as easily observable for apical and velar articulations, we believe that at least a crude descriptive system based on muscular effectors would have developed a long time ago. Unfortunately, lingual muscular events are not only hidden from easy viewing but are also complex in terms of the number of muscles involved and the manner in which articulators will be displaced for a given timing and activity pattern. Nevertheless, informed guesses can be made and then subjected to laboratory testing.[24] For example, the styloglossus is a bilaterally-paired lingual muscle whose fibers run from the tip back along the lateral margins of the tongue, finally exiting the tongue near the glossopalatal arch, where they curve upward to insert mainly into the styloid process of the temporal bone (Kaplan 1971:367). Its contraction both retracts and elevates the tongue, but in temporal coincidence with muscular activity which pulls the tongue forward and makes the tongue body rigid (genioglossus and intrinsic tongue muscles), its chief effect will be to raise the tip and margins of the tongue towards the alveolar ridge.[25] If we include glottal opening, we have a confluence of muscular activity or local articulatory maximum, which will serve as a partial description of the muscular events involved in what is traditionally transcribed as [t^h].

Now let us imagine what would happen if one of the component events of this local articulatory maximum, styloglossus activity, were to gradually reduce, while the remaining components remain constant. If the apex of the tongue was originally vectored to the alveolar ridge, a diminishing upward and backward force should allow the apex to move down and forward (alphabetically, an alveolar [t^h] would become a dental [t̪^{h}]). Continued styloglossus activity reduction would produce an interdental vector, transcribed as [θ]. Further erosion of this now-diminished local articulatory maximum could involve the remaining lingual events – the genioglossus and intrinsic tongue muscle activity. Although we are not prepared to speculate on the (probably multiple) possible intermediate vectors this further reduction would produce, their total reduction would leave only laryngeal opening, transcribed as [h].

It is important to remember that we are describing only a few of many possible reduction pathways. Exactly which pathway is taken over a long period of time will depend on the order, magnitude, and specific timing patterns of the muscular events associated with the local articulatory maximum undergoing reduction, not to mention temporally adjacent articulations whose effect we have disregarded for the sake of expository simplicity.

Of particular interest to us are those members of the strength/sonority hierarchy which are relatively weak. If we are correct in interpreting these items as residues of more complex local articulatory maxima, then their relatively minimal muscular content could serve as a rough indication of the set of events which actually constitute the "independently controllable aspects of the speech event" (Chomsky & Halle 1968:297). Unfortunately, liquids and glides, the weakest members of the hierarchy next to vowels, come nowhere close to an exhaustive list. Other obvious candidates include [ʔ] and [h], although they do not fit neatly into a 'low strength equals high sonority' schema, nor do their category labels – stop and fricative – reflect their articulatory weakness.

Some recognition of the reduction of articulatory substance in the more sonorant members of the strength hierarchy is to be found in Lass (1976:151ff.). However, in dividing the vocal tract into oral and laryngeal gestures, Lass leaves the impression that this spatial division is in some way significant in phonetic evolution. Unless some clear intra-group dependence can be demonstrated, we see no reason for treating the muscular events underlying laryngeal gestures as a group distinct from those which underlie oral gestures. Our default assumption is that the activity of an individual speech muscle is independent of all other speech muscles and blind to arbitrarily imposed spatial divisions. Where this assumption can be shown to be wrong, e.g. in reciprocal inhibition or gamma feedback loops, this stance will have to be modified. Until such evidence is forthcoming, coarse subdivisions of articulatory gestures such as those proposed by Lass (cf. also Halle 1983) are of questionable utility.[26] This criticism aside, Lass makes a major contribution in representing, for example, the frequent passage of [f,s,θ,x] to [h] as a loss of articulatory properties rather than the flipping of a few signs on features. It is only a short leap from this perspective to the realization that certain glides and liquids are similarly reduced in articulatory content and should be so represented relative to the local articulatory maxima which gave rise to them.

There is no clear dividing line between these weak consonantal articulations and those articulations transcriptionally rendered as vowel qualities or 'secondary articulations' of consonants. Acknowledging the existence, learnability, and relevance to diachrony of fine-grained substantive and temporal patterning allows us to regard as normal what for segmental and featural representation have been traditionally problematic cases – such as whether the [r] of American English *bird* in those dialects which retain it is to be rendered as a consonant or as a vowel quality; whether the [ʔ] of North Vietnamese 'broken-tone' (as in [kuʔu] (*cũ*, 'old')) is a consonant or part of the tonal quality of the vowel; whether the first sound in English *huge* is an independent [h], a voiceless palatal

glide, or a laryngeal fricative with secondary palatalization; and whether the initial articulations in Mexican Spanish *bueno* constitute a voiced labiovelar fricative, a velar fricative with secondary labialization, or a velarized bilabial fricative. Such questions are particularly difficult to answer in frameworks which fill segmental slots with items from a universal phonetic alphabet. In the non-segmental, non-featural conception which we advocate here, such difficulties do not arise.

It is sometimes observed that successive travel along the same decay pathways seems to be characteristic of the histories of particular languages or families, so that, for example, all (or particular) voiceless stops repeatedly reduce via voicing in some cases, via fricativization in others. Patterns of temporal reduction may also be seen as typical for a given language or family for centuries, leading us to speak of, for example, palatalizing languages. An important question here is how such language- or family-specific tendencies are best interpreted. Does it suggest that the processes are somehow 'chosen' by virtue of being familiar, or, rather, that what we are seeing is evidence that the seeds for descent along particular pathways are present at the outset – e.g. that the particular temporal and substantive muscular event patterns of voiceless stops in fact differ in voicing vs. fricativizing languages. If so, it suggests that language-specific differences constitute sets of alternative starting conditions which in effect are predictive of certain pathways of temporal and substantive reduction. If a particular configuration of muscular events is required for a given decay pathway, the absence of that pathway in a family may indicate that either the requisite event timing pattern, or some specific events themselves, are lacking.

5.4. Reductive processes under laboratory conditions

Current technology requires the use of electromyography (EMG) to record muscular events at the level of detail necessary to demonstrate temporal and substantive reduction across the formal-to-casual continuum where we have claimed these processes may best be observed. EMG is an invasive procedure: fine wire electrodes are implanted in the muscles of interest and the distal ends of these wires are connected to amplifiers to bring the signal up to a level compatible with either analog or digital recording equipment. A detailed description of our recording procedures and apparatus can be found in Mowrey & MacKay (1990), and we note here only that the subject (one of the authors) was seated in a comfortable chair and could move and speak freely during the data collection phase thanks to a lightweight source-attached preamplifier system.

5.4.1. Method

Electrodes were inserted in the left anterior belly of the digastric (ABD) muscle, 2cm posterior to the mental symphysis, 5mm lateral to the midline, to a depth of approximately 3mm. An additional pair of electrodes was inserted into the orbicularis oris inferior, but data from this insertion were found to be unsuitable for single motor unit analysis and were discarded. The ABD is very active during spontaneous speech, serving to lower the mandible especially (but not exclusively) for articulations involved in consonant to open vowel transitions. Our purpose in monitoring the ABD was to observe timing and magnitude changes associated with vowel reduction.

After electrode placement the subject was seated, connected to the amplification equipment and fitted with a microphone headset to insure a constant distance from the mouth to the microphone while allowing complete freedom of movement. The voice track and EMG signals were recorded on FM tape for subsequent playback and computer processing.

The first phase of the recording consisted of 25 minutes of spontaneous conversation. We felt that this was sufficient time to catch a few examples of vowel reduction in relatively unmonitored speech. No attempt was made to guide the conversation nor to elicit particular speech tokens. After completion of this phase, we played back the entire conversation, selecting candidate 'casual speech' sentences for use in the next phase.

In the second phase of the recording the subject listened to a selected sentence several times and then attempted to mimic the conversational speech token's rate, intensity, intonational contour and phonetic content as accurately as possible. Up to 20 imitation trials were recorded for each selected sentence with a pause of at least one second between trials. Immediately following each imitation set the subject attempted another set of repetitions, this time at a rate slightly slower than the imitation casual set, maintaining the intonational contour and intensity while avoiding hyperarticulation. Practice trials were performed until the subject felt that his output was sufficiently consistent. Up to 20 trials were then recorded.

The goal of this procedure was to produce two stylistic types, the first more casual and the second more formal, for each sentence selected from spontaneous conversation. In subsequent text we will refer to the two types simply as 'casual' and 'formal'. We are of course aware that these types are neither the most casual nor the most formal that the subject might produce, but we feel confident that they represent two valid exemplars along the casual-to-formal continuum. Earlier investigations of stylistic variation have required subjects to produce target utterances at a 'conversational' or 'normal' rate and then repeat these

utterances at a 'faster' rate. The faster rate is then implicitly or explicitly equated with a casual speech style. An example of such a study is Zsiga (1992), which also discusses the major problem with this methodology (p. 54): subjects may use different strategies to attain the faster rate, including a style which may best be characterized as 'fast-careful' speech. We do not expect fast-careful speech to exhibit the reduction processes commonly found in historical change and synchronic variation studies. We also assume that casual speech is faster than more formal styles, but we take the opposite approach to elicitation by requiring the subject to produce a slower version from a more casual model. We believe that in a laboratory setting subjects are more likely to be able to change articulatory styles from casual to formal than from formal to casual.

The repetitions performed for each style are necessary for the purposes of noise reduction through signal averaging, a standard procedure in EMG studies. Repetition is certainly not compatible with unmonitored speech production, one of the key characteristics of truly casual speech. Nevertheless, it would be quite surprising if imitations of casual speech differed substantially from spontaneous casual speech, and the acoustic characteristics of imitated vs. actual casual speech give no such indication, even to the trained ear.

5.4.2. Analysis

Repetition sets were digitized at a 20KHz sampling rate and stored on disk for analysis using a custom software package which allows visual inspection of all channels, playback of any portion of the voice channel, and signal averaging of selected time periods for the EMG channels. Each repetition was examined individually for EMG artifact and anomalous production (intensity significantly deviating from the norm or intonational contour not properly maintained). Repetitions failing this qualitative judgment (around 10% of the total) were excluded from subsequent analysis.

Three views of the EMG signal were derived from the data sets. (1) The raw EMG was integrated with a 30ms time constant. (2) The largest single motor unit (SMU) in the raw EMG signal was discriminated by a combination of computer software and visual confirmation as in Mowrey & MacKay (1990), an average action potential was created (aligned temporally on the point of peak action potential amplitude), and this average was digitally subtracted from the signal for every firing time of this unit. A new integration of the resulting signal was then carried out. (3) Firing times of the subtracted SMU were used to create a third view, with each firing time represented as a spike of unit amplitude.

EMG averages for both the casual and formal styles are on the same vertical scale, i.e. they have not been normalized. View (1) is the

standard for EMG studies and was used for the averages labelled in Fig. 4 as ABD+. Views (2) and (3) were created as a check on electrode placement, signal stability, and single muscle recording. View (2) assures us that the EMG signal is not disproportionately represented by a single unit, and that amplitude fluctuations present in View (1) persist in the absence of a single proximal unit. View (2) was used for averages labelled in Fig. 4 as ABD-. View (3) provides verification of single muscle recording by showing that the firing pattern of a single unit approximates the activity patterns observed in the gross EMG averages derived from Views (1) and (2). Traces labelled SMU in Fig. 4 are superimpositions of all repetition trials for each style.

For averaging alignment a single easily identifiable acoustic transition in the voice channel is selected. Whenever possible we choose a location near the middle of the utterance in order to minimize smearing of events more distant from the alignment point. The voice channel is also used to obtain average utterance lengths for the two styles.

5.4.3. Results

Here we present a single formal-to-casual pair to illustrate several points. The selected sentence, *I suppose that's true*, contains a reduced vowel in the first syllable of the word *suppose*. In the casual speech original and in 15 out of 17 casual speech repetitions, this vowel manifests itself in the voice channel as completely voiceless or with very low amplitude oscillations accompanied by frication noise. The remaining casual speech repetitions show a single glottal pulse at this location. Formal speech repetitions show the reverse pattern: one or two glottal pulses in 15 out of 18 tokens and voicelessness for the remaining. Representative trials for formal and casual styles are labelled in Fig. 4 as Voice(F) and Voice(C).

The alignment point for ABD averaging and SMU superimposition is indicated as zero on the time line for formal and casual averages. The alignment point selected was the [ð-æ] transition in *that's*. This point is characterized by cessation of high frequency noise and the onset of periodicity for the vowel. Fig. 4 shows that the onset of ABD activity precedes the onset of corresponding acoustic transitions by some 100ms. (The relatively early onset of ABD activity for the word *I* is due to the fact that in sentence initial position voicing does not begin for this word until the mandible achieves a lowered state.)

For the reduced vowel in *suppose*, ABD activity for the formal style shows a small but significant rise in the expected position. This is indicated by the arrow on the left over the ABD+(F) average in Fig. 4. For ABD+(C), the casual style shows a corresponding rise, but it is much closer to the major rise in activity associated with the [p^{h}ow] transition.

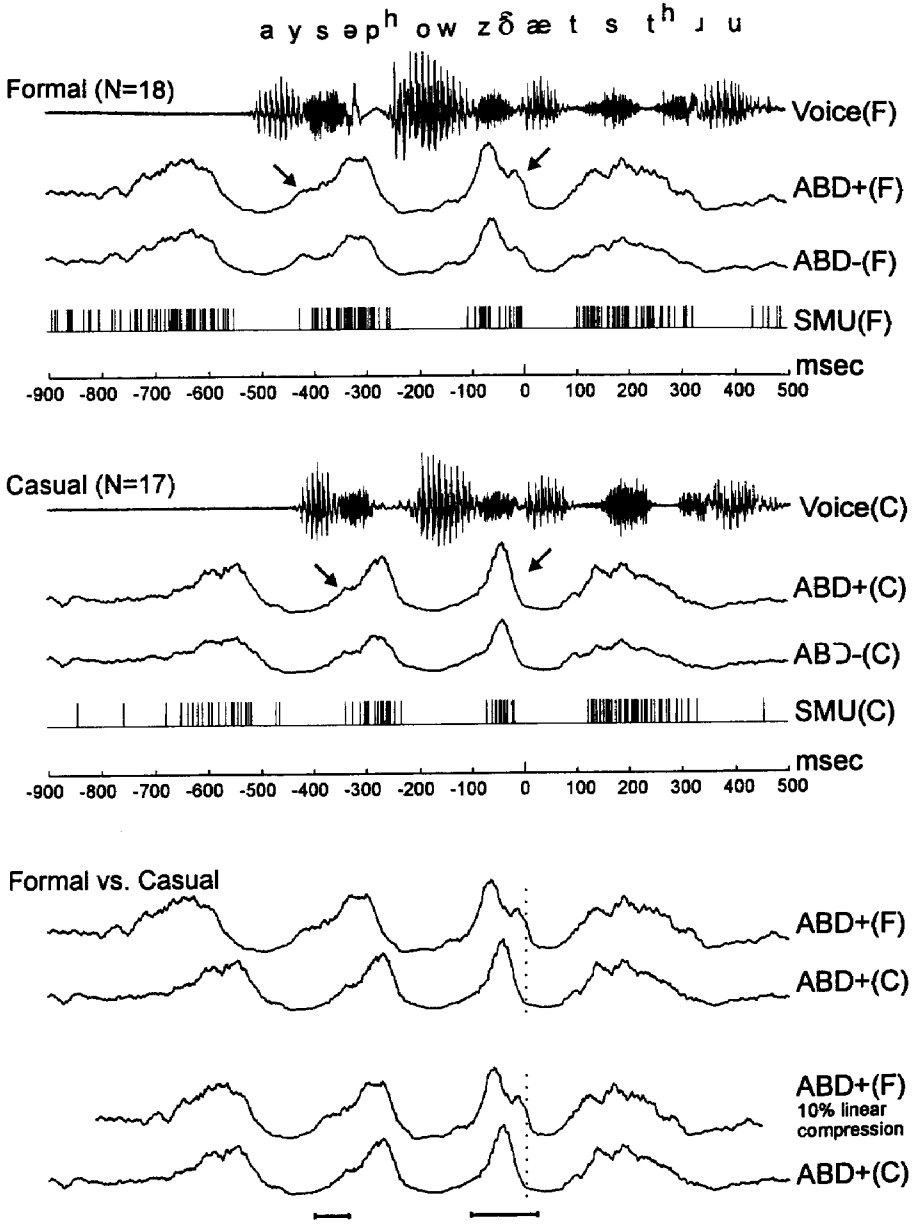

Figure 4.

It is important to note that ABD activity persists in the casual style in spite of the fact that the reduced vowel would not be noted in broad phonetic transcriptions of most of the casual style repetitions.

A second and unanticipated locus of reduction occurs for the articulation of *that's*. The ABD burst structure for this word appears as biphasic in the formal style averages ABD+(F) and ABD-(F) and two phases can also be seen in the SMU(F) superimposition. The first phase of this complex can readily be correlated with the [ð-æ] transition and was expected given the open nature of the vowel. The second lower amplitude phase (indicated by the arrow on the right in the ABD+(F) average) occurs at a time which would correlate with the [ts] transition using the 100ms lag time rule-of-thumb mentioned above. That an active muscular event associated with mandibular lowering occurs for this transition is interesting in its own right: one might assume that the [ts] transition is solely of lingual aetiology. In any case, the participation of such a muscular event is not unreasonable here as part of the release of [t]. A re-listening of the casual speech tokens reinforces this position. The sequence *that's true* in these tokens would be transcribed as [ðæstʰru] i.e. with a reduced or absent [t] in *that's*. The casual speech averages ABD+(C) and ABD-(C) show a single phase of activity for this sequence. Note that it is somewhat intermediate in time relative to the two phases of the formal style averages.

Averages presented under Formal vs. Casual in Fig. 4 show the time and amplitude characteristics of the two styles. The upper two traces are copies of ABD+(F) and ABD+(C) from above. In general, amplitudes of the muscular events in the casual style are slightly lower than the corresponding muscular events in the formal style. As expected from the design of the experiment, the temporal course of the casual style is shorter than that of the formal style. This is also evident in the example voice traces. As a first approximation, it might be suggested that the casual style is simply a temporally-compressed version of the formal. To test this we subjected the formal style average to a 10% linear compression, since the voice traces of the two styles differed in duration by about this amount across all tokens of the respective styles. The lower two traces show the formal style under 10% linear compression compared with the unaltered casual style. (Temporal alignment is shown by the vertical dotted lines.) Overall, the fit is quite good, but substantial differences at the reduction loci discussed above persist and are indicated by the horizontal bars at the bottom of the figure.

5.4.4. Discussion

In comparing the formal with the casual averages we make the following three observations. First, there appears to be no radical

reorganization of muscular events for the two styles examined. Components present in the casual style are also present in the formal style, and new muscular events are not introduced to compensate for the change in speaking rate or style. Although we cannot yet say with certainty that this holds true for the rest of the muscular events involved, it is at the very least a reasonable hypothesis. The simplicity of such changes in the motor program for articulation would not be expected if muscular events were enslaved to some higher level of planning where phonetic targets or spatially-defined gestures are transformed, deleted, or missed, and then sent off to some module for a do-as-you-must translation into any muscular events necessary to achieve the new spatial goals.

Secondly, masking (the hiding of gestures due to temporal overlap in fast speech), as proposed by Browman & Goldstein (e.g. 1990:363ff.), does not seem to be a necessary or sufficient condition for the formal-to-casual processes involved in either the vowel or the consonant reductions above. Masking implies maintenance of articulatory activity across styles with concomitant temporal compression, producing overlap. While we have no doubt that overlap of articulatory trajectories in a spatial coordinate system occurs due to tissue elasticity and inertial factors, temporal and substantive reduction of the muscular events underlying these trajectories exceeds the simple linear compression implied by this model.

Finally, we should note that the reduction in *I suppose* is an example of a family of semantically related reductions which demonstrate problems inherent in any theory of the compositional type, including the gestural theory of phonology. The members of the family are limited to *I* + verb collocations wherein the verb is semantically weak and expresses the speaker's volitional disposition or estimate of probability for the proposition which follows:

(12)	a.	I suppose	I suppose I should go.
	b.	I guess	I guess I'll go to the store.
	c.	I think	I think it's gonna rain.
	d.	I reckon	I reckon I'll mosey on down to the corral.
	e.	I bet	I bet you wished you hadn't said that.
	f.	I hope	I hope to see you soon.
	g.	I wish	I wish I was going with you.

All of these I + verb expressions have phonetic variants with highly reduced and often deleted first person pronouns, as in *Guess I'll go to the store*. In addition, some show clearly reduced articulations in the verb, as in [spowz] for *suppose*, and [hĩŋk] for *think*.

The theory proposed here treats, for example, the *I think* expression above as an indivisible entity, holistically stored and produced, and

distinct from ostensibly similar *I* + verb constructions, such as in *I think, therefore I am*. A compositional theory, on the other hand, by treating all *I* + verb constructions as composed of sequential building blocks, cannot distinguish between the constituent *I think* articulations in the following (both mundane) sentences, yet in the second sentence the *I think* does not reduce to *think* or [hĩŋk].

(13) a. I think it's gonna rain.
b. I think a lot before I make a decision.

This differential susceptibility to reduction is not predicted by any model which converts sequential building blocks to articulatory action sequences. On the contrary, compositional models such as gesture theory predict that the above *I think* constructs will be equally susceptible to reduction through gestural overlap and hiding when their gestural scores are run through the appropriate machinery.

5.5. Summary

The kinds of changes to muscular events which we expect in internal phonetic evolution and synchronic formal-to-casual processes are schematically represented in Fig. 5. The arrows indicate the directionality of the change or process, and our theory forbids changes in the opposite direction. The theory is testable (falsifiable) for synchronic formal-to-casual processes, albeit with some difficulty. For extension into the wider diachronic domain, we must appeal to the preponderance of such changes as obtained even through reconstructive methods which do not have temporal and substantive reduction as a theoretical cornerstone, and to logical argumentation.

Substantive and temporal reduction are complementary not only in that their respective dimensions are the two along which articulation must be defined (the muscular events and their temporal courses), but

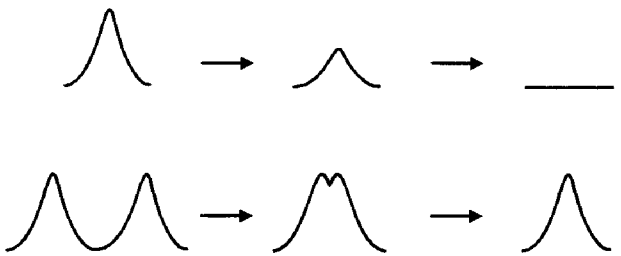

Figure 5.

also because it appears to be unlikely that there are any reductive events characterizable as either purely substantive or purely temporal. That is, substantive reduction seems to be accompanied by temporal reduction and temporal reduction by substantive reduction – or, to use more familiar terms: assimilations probably always entail some substantive reduction and lenitions some temporal reduction of surviving muscular events (Pagliuca & Mowrey 1987).

To see that it is not surprising that this should be so, we have only to consider the three logically possible ways in which substantive and temporal reduction may be related independently of particular lexical material, i.e. in terms of the articulation of expressions generally:

(14) a. substantive reduction exceeds temporal reduction
b. substantive reduction equals temporal reduction
c. temporal reduction exceeds substantive reduction

These three possible relationships can be restated in terms of mechanics (cf. Lindblom 1983) by pairing each relationship with its evolutionary consequence as follows:

(15)

In phonetic evolution, average articulatory energy expended per unit time	a. decreases b. remains constant c. increases

Of these three, (15a) must be ruled out, since it would eventually lead to collapse of the communication system, and we do not see languages which appear to be headed towards a catastrophic state of reduction-to-silence. (15c) is also problematic in the long run, since hyperarticulation would ensue if muscular events piled up on each other to ever greater degrees. Only (15b) accords with observation. Thus, only relationship (14b) allows for articulatory equilibrium under conditions of constant change and is consistent with the observed directionality of linguistic change.[27]

Our overview of substantive and temporal reduction now complete, we turn to an evaluation of the case for evolutionary epenthesis.

6. Non-epenthesis type 1: Transitional articulations

6.1. Oral stop excrescence

The historical developments exemplified in (16), whereby original nasal consonants appear as sequences of nasals followed by homorganic oral stops, are typically analyzed as instances of segment-insertions.

(16)	AGk anros	>	andros
	AGk mesemria	>	mesembria
	Lat numeru(m)	>	OF nombre
	Lat teneru(m)	>	OF tendre
	OE bræmel	>	ME bramble
	OE slumerian	>	ME slumber

Classical generative phonological description rendered phenomena such as those in (16) as formally distinct from assimilatory phenomena, such as regressive vowel nasalization, which required only adjustment in the featural content of existing segmental matrices. Recent generative treatments differ only superficially. Manipulation of a nasality autosegment (Wetzels 1985) or of metrical nodes (Piggott & Singh 1985) merely forestalls the inevitably necessary segmental reckoning: since mappings into tiers or nodes on which segments or segment positions are assumed to have their reality must be made explicit, even in such progressive versions of the theory one must still speak of C-insertion. A more recent version of such analyses is discussed in 6.4.

That changes such as those in (16) do not actually involve insertion has been argued for some time. Jespersen for example, commented on examples like OE *bræmel, slumerian*: "[the [b]] is due to the soft palate going up a moment too soon in the combination ml, mr" (Jespersen 1909:19).[28] Unlike the usual verbal characterization of such a development as an insertion of an additional consonant, this view recognizes that the temporal course of one component of the articulation of the nasal consonant has changed, the result most accurately describable as a partially-denasalized stop. Such changes thus arise from reductive alterations in the duration and relative timing of existing muscular events, much as in vowel nasalization and other assimilations; no new muscular events are introduced.

Partial denasalization of nasal stops is neither idiosyncratically Indo-European nor limited to pre-liquid environments, as the following additional examples make clear. Partial denasalization of the Proto-Northern Iroquoian nasal *n is evident from the Huron and Wyandot reflexes: [ⁿd], but [n] if a nasal vowel precedes or follows (Rudes 1976:17 et passim). According to Rudes, the nasal onset of the [ⁿd] reflex was often lost, leaving simple [d].[29] Partial denasalization of initial nasals before oral vowels in Korean is reported by Chen & Clumeck (1975), who outline a progression of the process from high to low vowels and give reason to suspect that denasalization will continue to completion, leaving oral stops. They also note that denasalization of nasal stops is recorded for certain Cantonese and South-Western Mandarin dialects of Chinese. n > nd, m > mb before oral vowels is reported for Indo-Iranian dialects by Gray (1902).[30]

Note that for partially denasalized stops to denasalize completely eventually, the events which open the velopharyngeal port must reduce substantively. Whether this reduction is also the sole event creating partially denasalized stops from nasals is not clear, although for nasals in initial position it is unlikely that temporal reduction – leftward displacement or anticipatory effects with no diminution of duration or extent of nasalization – is responsible for denasalization.

Partial denasalization of stops before [s] is reported for Catalan (ns > nts) (Brasington 1973; Roca 1975); Breton (ms > mps) (Lewis & Pedersen 1974); and before [s] and [ʃ] in English (ns > nts; nʃ > ntʃ in, e.g. *answer, ancient*) (Jespersen 1909; Dobson 1968; Zwicky 1972). In these cases, the glottal configuration of the following voiceless consonant is evident in the oral portion of the partially-denasalized stop, indicating changes in the relative timing of glottal as well as velopharyngeal events.

Complete denasalization of nasal stops is inferable as a historical process from comparative materials in many language families (in Penutian (m > b in the Lathrop dialect of Yokuts (Shafer 1947)); Athapascan (Hare (Hoijer 1966)); Nootkan (Nitinat (Haas 1969)); some Coast Salish dialects and Quileute (Hockett 1955:119); in various languages of the Eastern family of the East New Guinea Highland stock (Bee 1973); and elsewhere throughout the world). Correspondences of nasal stops with voiceless oral stops also suggest denasalization as a diachronic process, as do correspondences with various forms of r and l, the latter two as subsequent decays of denasalized stops, also prevalent in comparative materials. Allophonic and morphophonemic alternations of some of these sets in single languages are similarly interpretable as the result of temporal and substantive reduction.

6.2. Transitional stops by temporal reduction

The examples in (15) (Slavic handbooks, Andersen 1972; Lewis & Pedersen 1974; Dobson 1968, Zwicky 1972, Ohala 1974a, Javkin 1978) illustrate coarticulation effects involving lingual articulations, whereby an apical stop develops between a fricative and a liquid.

(17)	Slavic	sr > str, zr > zdr
	Celtic	sr > str > tr, sl > stl
	English	ls > lts, lθ > ltθ

As with the nasal processes in (16), alphabetism suggests that these phenomena involve the intrusion of segmental material. And, again, feature representations provide a less than optimal view. In this case, it is more than a failure to model-in temporal reference: consonantal

features, designed for efficiency in capturing contrasts, model lingual characteristics poorly. In particular, the location and extent of the tongue's contact with the palate in [s, r, l, t] have not been coded into features. Although the articulation of what is transcribed as [s] varies in detail from one speech community to the next, we may assume that for the data at hand it likely involved contact of the tongue along the sides of the palate, with either partial or uneven contact or approximation of the tongue apex with the anterior portion of the palate. Contact in [l] is apico-anterior only, the sides of the tongue being drawn in medially, away from lateral closure. The stop [t], regardless of the extent and precise location of anterior contact, is produced with complete closure at both anterior and lateral positions. The areas of contact for [s] and [l] are thus complementary; the stop exhibits both types. Although it is not precisely the case that a [t] 'is' an [s] plus an [l], their coarticulation at juncture close enough to result in coproduction of some of the muscular events of each may give rise to a sufficiently [t]-like configuration – an epenthetic stop.

The examples involving r are of special interest, as they permit us to infer the nature of the historical r on the basis of its participation in the development of a stop. The coarticulation analysis encourages us to posit as antecedent an apical r with sufficient anterior tongue-palate contact (and not, for example, an uvular r or mere retroflexion of the tongue). Such a diagnostic can be applied in a more predictive fashion as well. Zwicky (1972) discusses examples in English of both of the types of coarticulation-induced stops we have been considering. He provides examples of nasal and of [ls], [lθ], and [lð] contexts, and notes the non-participation of r in such variation. Thus he identifies the partially-denasalized stop of *dreamt, answer*, etc. and the transitional apical stop of *pulse, Welsh, health*, as casual-speech objects, but cites as unlikely the rise of transitional stops in such words as *course, harsh* and *cars* (p. 291ff.). The temporal reduction analysis allows us to do more than note non-occurrence. We can predict that, given the generally weak character of such r's in American English, in which contact is often not involved, r cannot provide the complementary closure necessary for the formation of a transitional [t] through coarticulation with [s]-closure.

Of similar aetiology is a development in Baule, a Kwa language, cited in Vago (1976). Baule [l] is realized as a flap when it occurs following an alveolar or palatal consonant, but remains [l] when following a labial or velar. Vago interpreted this as evidence for the acoustic feature [grave], as no evident articulatory analysis could explain such behavior. As we have argued elsewhere (Pagliuca & Mowrey 1980), this change has nothing to do with the acoustic properties of the consonants involved, but represents a coarticulation effect much like the cases we

have been discussing. The result of coarticulation in Baule, however, does not give rise to a new or inserted segment, and this alphabetic difference obscures the similarity: partial temporal coincidence of [l] with an alveolar or palatal stop or fricative provides complete closure, a short-duration [t]-like transition or flap, much as complete closure at the juncture of [l] and a sibilant in the Indo-European examples of (17) gives rise to a transitional stop.[31]

Temporal reduction of existing muscular events also gives rise to the [t]-like transition or flap in the American English dialectal forms in (18), cited by Lass (1984) as a fortition. Here, some lingual events in [z] have reduced, permitting a somewhat earlier closure for [n], resulting in a brief non-nasal transitional flap.

(18)	*isn't*	*doesn't*	*wasn't*	*business*
	['ɪDn̩]	['dəDn̩]	['wəDn̩]	['bɪDnɪs]

6.3. *Indications of the reductive origin of transitional stops*

Some evidence relevant to our claim that temporal and substantive reduction are typically copresent in evolutionary changes is available for some of the change types we have just discussed. Fourakis & Port (1986), measuring durations of the components of [ls], [lts], [ns] and [nts] sequences in South African and American speakers of English from spectrograms, showed that the transitional [t]'s of items such as *dense* and *false* in American English are significantly shorter than the underlying [t]'s in *dents* and *faults*. They also note that the nasals in words with transitional stops (*dense, tense*) are longer than those in words with underlying stops (*dents, tents*), which they interpret as a direct result of the transitional stop. This interpretation, however, seems to be incorrect. The length difference in nasals does not appear to be a consequence of the formation of the transitional stop: nasals are also significantly longer in [ns] sequences than they are in underlying [nts] sequences in South African English, in which [ns] sequences never give rise to transitional stops. Thus the transitional stops cannot be regarded as adding to the temporal duration of either the nasals or the words in which they arise. For all relevant pairs involving nasals, there was no significant difference in overall word length, measured by mean durations in ms. over all tokens. Since *dense* does not arise from *dents*, some other basis for comparing forms with transitional stops to their likely immediate precursors in the same dialect is required.

For the South African subjects, for whom word length differences in these pairs are also not significant, both the nasal *and the fricative* are significantly longer in [ns] over [nts] sequences. This suggests a way of

estimating the likely durations of nasals and fricatives in American underlying [ns] sequences as they were prior to the creation of transitional stops: extrapolate their values from the differences between the mean durations of nasals and fricatives in [ns] vs [nts] sequences in South African speakers and the observed duration of nasals and fricatives in American speakers in underlying [nts] sequences by treating the differences as proportions. Thus, solving for x (as in (19)) for each nasal and fricative in each [ns]/[nts] pair, where x is the expected (prior; original) duration, provides the expected durations for nasals and fricatives in original American [ns] sequences; (19), for example, gives 71ms as the value for the expected duration of the nasal in Amer. *dense* ($61x = 4350$; $x = 71$). Although this is obviously an inexact method, the sum of the differences between the expected and observed durations of the nasal and fricative gives a reasonably close match with the observed duration of the transitional stop; cf. Table 3.

(19)
$$\frac{\text{observed duration of nasal in S.A. } \textit{dense} \text{ (87)}}{\text{observed duration of nasal in S.A. } \textit{dents} \text{ (61)}} = \frac{\text{expected duration of nasal in Amer. } \textit{dense } (x)}{\text{observed duration of nasal in Amer. } \textit{dents} \text{ (50)}}$$

Table 3.

	Nasal duration		*Fricative duration*		*Sum of duration differences*	*Observed duration of stop*
	Expected	*Observed*	*Expected*	*Observed*		
dense	*71*	*58*	*120*	*103*	*30*	*31*
tense	*69*	*54*	*110*	*98*	*27*	*36*

For ls/lts test items (e.g. *false* vs. *faults*) [l] was not measured directly; rather, the syllabic nucleus of all words with [l] was measured from 'the release of the closure of the initial consonant to the point where all the [l] formants disappeared before [t,d] or the onset of high frequency noise for [s,z] as a single unit' (p. 203). Here, too, there is a close fit between the sum of the differences between expected and observed values for the durations of the nucleus and fricative and the observed duration of the transitional stop for the American subjects; cf. Table 4.

Tables 3 and 4 suggest, if only indirectly, that the transitional stops arise as products of changes in the temporal sequencing of the original

Table 4.

	Nasal duration		*Fricative duration*		*Sum of duration differences*	*Observed duration of stop*
	Expected	*Observed*	*Expected*	*Observed*		
false	*208*	*190*	*118*	*103*	*33*	*39*

nasal or liquid and the final fricative, which are thereby also substantively reduced.

6.4. Evaluating a phonological account of the phenomena

Clements (1987), arguing for a phonological treatment of the phenomena in (16) and (17), divides them into two classes in a very different way. Following Wetzels (1985), he classifies stops intruded before obstruents as type A and those intruded before sonorants as type B. As a result, type A and type B each include some instances of oral stop excrescence and some of temporal reduction. The temporal reduction sl > stl, for example, is grouped with nr > ndr in type B because [l] and [r] are sonorants rather than obstruents, whereas ls > lts, a temporal reduction differing from sl > stl only in the order in which the articulations occur, is type A, as is ns > nts, which we regard, along with nr > ndr, as an instance of oral stop excrescence.

Clements proposes that type A phenomena result from the lag of the oral occlusion of the nasal or liquid into the following obstruent rather than from early raising of the velum (p. 38), and are thus representable as the spreading of the oral occlusion value [-continuant]. It is unclear how the oral occlusion analysis, which assumes a following obstruent, can account for n > ^{n}d/ _V, though he cites this as a process which an adequate analysis must be able to handle. Moreover, reduction in velopharyngeal activity (palatoglossus activity and reciprocal levator palatine inhibition) is also required to explain the subsequent complete denasalization of partially denasalized stops; neither early velum raising nor the lag alternative, which attends only to the fact that oral occlusion is no longer temporally coextensive with nasalization, allows us to view n > ^{n}d > d as a single evolutionary decay series.

In part because he includes in B the changes lr > ldr, nl > ŋgl, and sl > skl, none of which he regards as phonetically motivated, Clements does not propose a spreading analysis for type B. Instead, to explain all type B phenomena, Clements appeals to a complex set of explanatory factors, including a 'remedial' strategy for improving syllable contacts

with rising sonority profiles and a 'substitution' strategy for avoiding ill-formed clusters. In type B phenomena

> [...] the first member of the host cluster is a fricative, nasal, or liquid and determines the place of articulation of the intrusive stop, as well as its voicing. The second member contributes the features [-nasal, -lateral] [...] Finally, with occasional exceptions, the intrusive stop always forms a possible syllable-initial cluster with the following consonant. (p. 40)

The exception Clements cites is Ger. *hoffentlich* (< *hoffen* + *lich*), [tl] not being a possible syllable-initial cluster.[32] Initial zl > zdl, reported for a dialect of the Italian Marches (Rohlfs 1966:263), which produces an initial cluster not previously found, is also an exception. More problematic for Clements' analysis is that Rohlfs also reports that dialects of the Marches generally exhibit zr > zdr (as do Neapolitan, Calabrian and Sicilian dialects). What allows Clements to group [ls] with [ns] and [lr] with [nr] is the assignment of [-continuant] to both laterals and nasals, but not [r]; this provides the required [-continuant] value for the intrusive stop. But where [l] and [r] are in contrast and zl > zdl and zr > zdr both occur, the analysis incorrectly predicts that only [zl] and not [zr] sequences will give rise to a stop, whereas zr > zdr appears to be more common here than zl > zdl. The analysis could be salvaged by assigning [-continuant] to [r] as well as [l], provided they behave alike with respect to other processes that are sensitive to the value for [continuant]. Insufficient information prevents us from knowing whether this is an option in the dialect cited by Rohlfs, but uncooperative [r]'s and [l]'s will make salvage impossible, in type A as well as type B cases. In the Basque dialect of Getxo, for example, stops arose in all [rs], [ls], and [ns] sequences in native words and integrated borrowings (Hualde & Bilbao (1992:3ff.), where the process is referred to as affrication of [s] following [r], [l] and [n]); otherwise, however, [r] behaves as a [+continuant], [l] as a [-continuant]: as in Spanish, the continuant allophone of /d/ – [ð] – appears after an [r], the stop allophone – [d] – after an [l].

These difficulties with Clements' analysis suggest that the type A/ type B division is neither as economical nor revealing as he supposes, and that, for diachronic purposes at least, a more straightforward account is provided by an analysis which assumes the two reductive phenomena of partial denasalization and temporal reduction.[33]

6.5. Transitional fricatives and glides

If coarticulation changes can give rise to transitional stops, we should not be surprised to find similar effects resulting in weaker

consonantal transitions. In fact, we might expect that the historical record evidence at least as many instances of apparently intrusive fricatives and glides as stops. But quantification is difficult here, and it is not clear that such cases outnumber those of stops. One reason for this may be that alphabetic traditions may not as easily record the consonantal products of coarticulations when they are relatively weak. The fact that the instances of fricative and glide insertions typically cited are described in more or less phonetic accounts rather than on the basis of orthographic evidence lends weight to such an interpretation.

Millardet (1911) documents a series of fricative consonant insertions in three Swedish speakers. These consonant insertions go well beyond the standard description of Swedish 'inrounded' and high front vowels as diphthongal in nature, sometimes transcribed as having postvocalic [β] and [j] offsets respectively.[34] Although some of what Millardet describes as fricative consonant insertions would be described as non-segmental transitional phenomena under current practices, an important residue – fricative insertions before voiceless stops – deserves careful attention: our theory does not allow local articulatory maxima such as fricative articulations to simply come into existence without the requisite complement of muscular events as precursors.

Briefly stated, Millardet found that corresponding to Standard Swedish vowel + voiceless stop sequences, his subjects produced vowel + fricative + voiceless stop sequences. He noted that the nature of the epenthetic fricative depended on the nature of the preceding vowel: back vowels give velar fricatives, palatal vowels give palatal fricatives, labial vowels give labial fricatives; cf. (20). Now the conditioning factors of this so-called epenthesis are phonetically quite plausible.[35] The question is, does it make sense to say that a fricative articulation has arisen here where nothing existed before?

(20)

V (pal)	+	stop [-voice]	>	V (pal)	+	fricative (pal)	+	stop [-voice]
V (lab)	+	stop [-voice]	>	V (lab)	+	fricative (lab)	+	stop [-voice]
V [back]	+	stop [-voice]	>	V [back]	+	fricative (vel)	+	stop [-voice]

Recall regressive vowel nasalization. Viewed as a process, vowel nasalization first appears as nasality over the latter portion of the vowel; nasality is a remnant articulation of a decaying consonant. Note that at this stage, we never speak of epenthetic nasal vowel insertion. In Swedish, it is the laryngeal state of the final voiceless consonant which

is encroaching on the latter portion of the vowel (recall the analysis of x > f in 5.1), but here, rather than viewing this as voicelessness retiming onto the vowel, phonologists insist on epenthetic segmental status for the resulting coarticulation. This inconsistent treatment of essentially identical processes is the result of a notational deficiency, and does not provide evidence of epenthesis as a language-internal development.

Glide insertions are also amenable to the sorts of analysis we have suggested. The diphthongization in (21), for example, which is analyzed by Lass (1984) as an insertion of a short [i] between non-high vowels preceding palatals and velars, is the result of the temporal coincidence of the doming tongue for such consonants with the latter part of the vowel; no new articulations are intruded.

(21)	Amer. Eng. dial.	*bag*	*cash*	*bush*
		[be^{i}g]	[ke^{i}ʃ]	[bʊ^{i}ʃ]

A similar effect is noted by Catford (1977:226), citing Gairdner (1925), as one of two ways in which the contiguity of palatal vowels and uvular stops in Arabic is 'accommodated' (Catford's term): either a glide with the character of a schwa or an unrounded high back vowel develops between the [q] and the vowel, or the uvular becomes 'somewhat centralized'.

6.6. Summary

The cases we have examined are often cited as prima facie examples of epenthesis and hence of the augmentative nature of at least some sound changes. Classical generative notation was incapable of representing the events presumably involved as anything other than insertions of entire feature complexes. Despite improvements in current versions of phonological theory, such changes are still analyzed as involving events which are substantively or temporally augmentative. As we have seen, however, the intrusive status of transitional stops, fricatives and glides is a notational fiction, arising from the serial alphabetizing of articulations whose time domain cannot and should not be so represented. Such phenomena are properly not insertions – no new material is intruded – but the result of the temporal and substantive reduction of existing muscular events.

7. Non-epenthesis type 2: Analytic epenthesis

It is clear that some apparent augmentations have their source in the notions of elegance and simplicity which guide synchronic and

reconstructive analytic procedure. Over-reliance on such notions only serves to add spurious changes to the catalogue of diachronic events. Under current practices, the choice between analyzing a given language-internal distribution in terms of vowel epenthesis in a certain environment or vowel elision (reduction to zero) in the complementary environment will be dictated by the number of environments the process encompasses and the completeness of the historical attestation. But because the choice may actually be controlled by the relative simplicity with which each of the two complementary environments is formally statable, analyses based on environmental distribution cannot be relied upon to reflect history. Thus, given a language whose history is unknown to us, if a deletion process had extended its domain to include all environments save one, and that one were simply statable, the temptation to posit an epenthesis would be strong. This is analogous to the treatment of other, superficially less radical reductions. For example, given a stop in some environment (E1) alternating with a fricative in a complementary environment (E2), the choice of an underlying form or protoform is at least in part determined by the simplicity with which a rule can be stated. Thus, despite the general awareness of the overwhelming incidence of lenition relative to fortition, a fricative will be chosen as original or underlying if E1 is more simply statable than E2. Here again, local distribution and descriptive elegance may foster analyses at odds with history and phonetic probability.

It is difficult to assess the extent to which such analytically-derived augmentations have contributed to distorting the diachronic picture, but it is likely considerable, especially since the predilection for simplicity and elegance in generative rule-writing greatly reinforced structuralist notions of economy and pattern congruity. Thus, the synchronic analysis of English regular plural-formation typically involves a basic allomorph {-z} with one rule to devoice it following stem-final voiceless obstruents, ordered after a second rule which inserts a schwa if the stem ends in a sibilant – the reverse of the historical reality, which was the erosion of the weak vowel in all environments *except* those in which the stem was sibilant-final. In Yip (1987), *all* old weak morphemic vowels in English are denied underlying existence and are instead derived by general rules; a similar analysis for German is found in Lieber (1987).

Whatever their merits in synchronic classificatory description, such analyses confer an unwarranted legitimacy on diachronic rules of epenthesis. Removed as they are from articulation and its substantive and temporal detail, the primitives of this tradition – classificatory features and alphabetic units – are inherently non-dynamic, as are rules, which are formal objects describing distributional facts. Rules may be adequate as devices in a formal system, operating on classificatory

elements, but since they do not necessarily reflect the evolutionary processes which gave rise to the distributions they are unreliable as guides to history.[36]

7.1. Epenthesis analyses based on comparison of dialects

Suppose we find five sister dialects displaying the distribution in (22) in some class of cognate lexical items.

(22)	Dialect A	Dialect B	Dialect C	Dialect D	Dialect E
	-CC-	-CC-	-CC-	-CC-	-CVC-

Given such a distribution, the temptation to reconstruct *-CC- in the parent lexemes is strong, and the reasons are obvious: first, the -CC- pattern is statistically heavily predominant; second, if we reconstruct *-CVC- instead, then, given no other changes suggesting a closer relationship between A B C and D as against E, we must assert that in four sisters there arose independently a reduction process which eventually eliminated the vowel. That there could have arisen four separate parallel developments (not necessarily temporally coincident) is regarded as less likely than a scenario involving only one event – an epenthesis in E. Clearly, the method is not insensitive to considerations of plausibility and economy. The question, however, is whether these considerations properly enter at this point, or are more appropriately and profitably introduced elsewhere. What we are suggesting is that, despite the apparently sound basis for such a decision, the cost it entails in this case and generally is too high: these criteria in effect provide us with a solution which runs counter to our expectations based on observations of evolutionary change generally. To see this, it is only necessary to consider how differently we would analyze the structurally very similar situation in (23).

(23)	Dialect A	Dialect B	Dialect C	Dialect D	Dialect E
	-V̆s-	-V̆s-	-V̆s-	-V̆s-	-VNs-

Confronted with the forms in (23), few would hesitate in reconstructing *-VNs- as the parent sequence, even though this 3-segment stretch occurs only in E, and all remaining sisters exhibit the 2-segment string -V̆s-. We would be no less confident of the correctness of our choice in this case than we were in choosing *-CC- for (22), but our decision is informed in a very different way. We do not choose *-VNs- in (23) on the basis of local statistical distribution, as we chose *-CC- in (22), but rather on the basis of universal substantive considerations: the knowledge that

the phonetic likelihood of *VNs > V̆s is great, and that of V̆s > VNs small guides our reconstruction. We possess repeated reliable observations of events of the former type, for which we can offer phonetic explanations, but neither observation nor plausible explanation are available to corroborate the latter.

We have discussed two situations in which epenthesis may be proposed, the first essentially equivalent to internal reconstruction, the second involving comparative reconstruction. In each case, allowing a local distributional pattern to favor epenthetic solutions is methodologically the opposite of what we do when faced with similar distributional patterns arising from well-understood global processes such as vowel nasalization. Since vowel reduction to eventual loss is no less of a globally statistically predominant process, we suggest that local distributions should be understood in terms of this process, and that epenthetic solutions based on distribution should be avoided.[37]

7.2. Phonetic explanations for vowel epenthesis

Since few analysts are content to propose epenthesis without offering some underlying motivation, phonetic explanations adduced for epenthesis must also be critically examined. The most common reason given for vowel epenthesis is to break up consonant clusters by inserting a vowel either between two consonants or before an initial cluster in order to realign syllable boundaries. In 4.2 we discussed the merits of such explanations for foreign vocabulary borrowed into a target language. As an explanation for internal epenthesis however, such analyses suffer from logical problems.

If we claim that a language breaks up clusters by epenthesis, we tacitly admit that it lived with these clusters for some period of time before the break-up, albeit in a state of sin. If the clusters are very old, we may wonder why the language waited so long to break them up. If the clusters are new (having arisen from vowel reduction, for example), we may question why the language allowed them to arise in the first place, as if seduced by the prospect of illicit consonantal conjunction. Sinking deeper into this quicksand, we may offer 'ease of articulation' or 'preferred syllable structure' as grounds for epenthesis. But the same questions come back to haunt us: why were 'difficult' clusters and 'bad' syllable structures allowed for so long if they are old, and, if new, why would a language move towards more difficult articulatory sequences and bad syllable structure? Since they can be suspended or applied at will, depending on whether consonant clusters are newly forming or are being broken up, explanations of this kind are not very helpful.

7.3. The authority of attested dialects

A third source of potentially spurious epentheses and other changes is traceable to the tradition of regarding as authoritative a single, earlier attested form of a language to the extent of assuming that extant modern dialects are necessarily its direct descendants. The more reasonable default assumption would appear to be that the language reflected in the literary monuments of a given period is but one of many contemporaneous dialects. The probability that in a given subfamily all modern dialects are direct descendants of one or more older, attested dialects is extremely remote; any number of the modern dialects may be descendants of unrecorded sisters whose phonetic characteristics may have been radically different.

It is not unusual to find historical cases of epenthesis posited on the basis of comparative evidence interpreted on the assumption of lineal descent. If we would only pause to evaluate the phonetic plausibility of the resulting scenarios based on what we know about change, our reconstructions might look very different, and they need not involve our compromising the genuine benefits of the comparative method. In this connection, consider the Oscan forms in (24), which Buck (1904:51f.) cites as cases of anaptyxis between a C and a following liquid.

(24)

patereí	(cf. Lat. patri)
pústiris	(cf. Lat. posterius < *postrios)

Given that the relevant paradigm for 'father', for example, must have originally been built on the stem form *pater-* by suffixation, is it not at least as plausible to assume that Oscan, rather than having first deleted and later reinserted it, simply retained the second vowel? One wonders whether, were the situation reversed, with Oscan the language of the empire-builders, and Latin attested only by ancient inscriptions, the reconstruction of the parent Italic dialect would not include the vowels present in Oscan.[38]

7.4. Parasitic stops

There are in the literature, however, claims of true consonantal epenthesis, that is, cases in which a consonantal articulation is thought to arise from either a weaker vocalic articulation or from no prior articulation at all. Perhaps the most widely discussed example is Burling's (1966) proposal of epenthetic final stops in Maru, a language of Northern Burma in the Lolo-Burmese subgroup of Tibeto-Burman (see also Burling 1967). Of the six Lolo-Burmese languages Burling examined, only Maru shows syllable-final stops for a given set of cognates; cf. (25).

(25)

Proto-Burmese	Burmese	Atsi	Maru
*ei	-ei	-i	-it
*o	-ou	-au (<*yo > ui)	-uk

Burling's argument for epenthesis rather than retention only in Maru centers around his reluctance to assign more than two contrasting tones to the stopped syllables of the reconstructed Lolo-Burmese ancestor. His goal is therefore to simplify the tonal system of the parent language by adducing consonantal epenthesis in one of the daughters. But this increase in simplicity is accompanied by a decrease in plausibility. First of all, Tibeto-Burman, as Sino-Tibetan languages generally, have witnessed a wholesale decimation of syllable-final consonants throughout their recorded and reconstructed histories. Many of these languages now allow only remnant glottal stops in syllable-final position. Is Maru, which participates in the trend toward reduction so prevalent in Sino-Tibetan and other language families, reversing it in this case? Secondly, Letsi, a language closely related to Maru, which Burling did not consider, also shows final consonants in some of the cognate items (Shafer 1974:339 et passim). Although we might interpret this as support for the insertion analysis, pointing to the possibility that Lolo-Burmese evidences two cases of insertion rather than one, it can just as easily be read as suggesting that the stops are relics rather than the products of innovation. Finally, Miller (1968, quoted by Benedict (1972:60); and p.c. to Matisoff (Matisoff 1968:882)) provides ancient Chinese cognates with final consonants in order to discredit Burling's proposal. Many of these cognates are in turn disputed by Benedict (1972); Matisoff, too, is unconvinced by Miller's analysis, accepting Burling's interpretation as a matter of course (Matisoff 1990:546). The point is that this proposed epenthesis is far from the clear-cut counterexample required to refute the proposal that consonantal epenthesis does not occur as an evolutionary development.[39]

Parasitic stops after high vowels are also reported for isolated Romance and Germanic dialects (e.g. Jeanjaquet 1931:45). For several reasons, such cases are particularly difficult to evaluate. First, there is the problem of the disappearance of these dialects in modern times as the isolation of their speech communities has diminished. Second, isolated dialects are typically found along geographical features which, although difficult to penetrate, often coincide with national and linguistic boundaries. But since boundary areas are the most prone to contention and invasion, they are also among the most likely to exhibit features of language contact. In effect, the linguistic advantage of geographical isolation is negated by the probability of infrequent but radically shifting linguistic dominance (see e.g. Tappolet (1931) for the

history of the dialect area discussed by Jeanjaquet). These two problems conspire in such a way that we cannot know whether to treat unusual features of such dialects as the result of language contact or substrata, as innovations arising and surviving because of isolation, or as archaic relics of the language family to which the isolate belongs. Finally, there is again the problem of data quality. Nineteenth and early twentieth century fieldworkers, however thorough, did not always have the means to transcribe the more subtle laryngeal articulations such as the glottal stop. With this handicap, it may not have been uncommon for combinations of high vowel plus glottal stop – [i?], [u?] – to be transcriptionally rendered as [it], [ik], [ut], [uk].[40]

To the extent that the dialects in question have disappeared and early transcriptions are of questionable accuracy, we must view with suspicion claims of parasitic consonantism. In support of the view that they can arise through evolutionary processes, evidence for such phenomena should be observable in synchronic variation studies as *emerging* parasitism, i.e. parasitic consonants beginning to appear and spread in the casual style of young speakers. We are aware of no such evidence. On the contrary, in those dialects where parasitic consonantism is given detailed attention, these articulations (whatever their origin and nature) are described as very attenuated and well on their way to complete loss (e.g. Jeanjaquet 1931:46ff.).

8. *Conclusion*

We have proposed the following general hypothesis. Internal or evolutionary change (sound change) consists of just two types of changes in muscular activity patterns: temporal reduction, by which bursts of muscular activity become temporally more contiguous, and substantive reduction, by which bursts of muscular activity diminish in amplitude. These two event types appear to provide straightforward accounts of most of the major types of phonological processes. In an attempt to demonstrate that the scope of substantive and temporal reduction is even broader, we have also focussed on the ostensibly disconfirming class of events to which epenthesis belongs. Evaluating the case for evolutionary epenthesis, we have concluded that it is weak. Some will no doubt feel that in using schematics and isolated examples to illustrate our point, we have sidestepped the genuine difficulties one encounters in trying to reconstruct a detailed and coherent history for a given family. But comparativists themselves point out that, confronted with distributions of just the type we have discussed, it is often difficult to know whether epenthesis or retention should be posited (e.g. Shevelov

1965:448f. for Slavic; Gray 1934:26 for Semitic). Our evaluation was prompted by the conviction that if we incorporate the strong hypothesis which the evidence for unidirectionality in internal change suggests directly into theory and practice, the way is open to a more coherent and accountable theory of change. The evidence for reduction is clear, repeatedly confirmed and supported phonetically; putative instances of epentheses and other augmentations should meet the same standards of data quality. But such data is typically of low quality, the evidence tending to be inferential rather than observational. We do not expect that all cases of internal epenthesis will be explicable, nor do we intend to explain them all away. But even if every case were valid, epenthesis would still be statistically insignificant relative to reduction, and does not seem to be observed in contemporary internal changes-in-progress.

Finally, we may speculate on the reason for the unidirectional, reductive character of change. The long-term diachronic trend is towards what might be viewed as signaling efficiency; whether or not it represents an increase in the density of information per unit time or unit energy, it must have figured in the emergence of human language from less densely-packed pre-human communication systems, and perhaps may be just barely visible within the short history of modern language. We realize that this may seem far-fetched to linguists, who tend to conceive of language in psychological rather than biological terms, but the ritualization and decay of displays in non-human communication systems studied by ethologists (cf. e.g. Moynihan 1970) are strikingly clear analogues of the grammaticalization and inexorable articulatory reduction which characterize the history of lexical material in human languages.[41] Signaling efficiency is likely directly tied to the evolutionary success of the species, conferring selective advantage; it is not epiphenomenal. The reduction of redundancy in signals or messages, especially those repeated often and therefore in part predictable, is automatic and advantageous. We see this most clearly in the relation of casual to more explicit formal styles of speech, which mirrors and foreshadows diachrony. It is an error to conceive of this as deriving from ease of articulation or laziness; such parochializations have detained us too long, preventing us from considering that use-induced reduction over time is characteristic of all symbolic motor behavior.

Addresses of the Authors:

Richard Mowrey: Human Neurosciences Research Unit, University of Ottawa, 451 Smyth Road, Ottawa, Ontario K1H8M5, Canada
William Pagliuca: Graduate College, University of Illinois at Urbana-Champaign, 108 Observatory, Urbana, Illinois 61801, USA

Notes

* Earlier versions of some of the material below appeared in Pagliuca (1982), Mowrey & Pagliuca (1980), Pagliuca & Mowrey (1987), and in the first version of the present paper in 1988. Our names here appear in alphabetical order; we have contributed equally to, and are equally responsible for, all that follows. We are very grateful to Carol Fowler for pointing out deficiencies in the first draft of this version of the paper and for making suggestions for improvements.

[1] Hockett (1965) cites the quantization (phonemic) hypothesis as one of the four breakthroughs of modern linguistics, and regards it as having had profound consequences not only for the conception of synchrony, but for its relation to diachrony as well.

[2] The generativist position here seems more to resemble than depart from the structuralist norm: in 'grammar change is rule change' we are not far from the implicit disdain for phonetic-level phenomena embodied in the structuralist precept that, unless a sound change affects the system, nothing (significant) really changes, i.e. if the system isn't thereby changed, neither is the language. Although, as Weinreich, Labov & Herzog point out (1968: 141f.), distinctive features allow changes which do not alter distribution to be recorded as genuine (rather than as irrelevant changes in phonetic realization), it is nevertheless the case that until a change has proceeded far enough to warrant a change in distinctive feature representation (and far enough through the lexicon to affect a majority of items) – i.e. until it has attained sufficient regularity for a rule describing it to be writable – it is not accorded recognition, and is thus denied grammatical status.

[3] Phillips (1984) suggests that changes which affect the most frequent words first are physiologically motivated, whereas those that affect the least frequent items first are not. Examining three of the latter sort – glide deletion in Southern American English, the unrounding of /ö(:)/ in Middle English, and the ongoing rise of diatones in English – she argues that their motivations are essentially psychological, and identifies analogical pressure and underlying segmental and sequential constraints as the relevant factors.

[4] In historical terms, its most reduced variant seems to be *Hi* (or *Hey*, depending on dialect) which, like *'Bye* (< *God be with ye*), is a one-syllable remnant of a complete sentence. In questioning whether it is reasonable to account for all the reduced variants of highly frequent expressions (such as *How's that?, What are you doing?, What's up?, I don't know* and so on) by starting from full forms, one may well be brought to the conclusion that what we are seeing here is evidence that what is acquired and stored are batches of register-dependent routines. Even for less frequent (and less sentence-like) expressions, multiple stored routines might make more sense; it is not clear, for example, that we would want to derive *gonna* and *lemme* from *going to* and *let me* each time by on-line processes.

[5] The psychological validity of the phonemic principle has been called into question on other grounds (cf. e.g. Liberman et al. 1974, Morais et al. 1979, Morais 1985, Warren 1976, 1982); such challenges have not gone unanswered, however (cf. e.g. Studdert-Kennedy 1987). Some instrumental evidence that speech errors (generally regarded as providing the best evidence for the segmental organization of speech), may not support traditional assumptions about segments is reported in Mowrey & MacKay (1990). For all its merits, the structuralist tradition has obscured the primary fact that becoming a native speaker involves mastering the fine articulatory details of a particular dialect (or multiple dialects, since individuals who in childhood moved from dialect area to dialect area may retain dialect-specific pronunciations for certain lexical items acquired in each dialect) and that dialects need not be geographically nor socially far apart for measurable differences to be discernible. As we have indicated, at least some articulatory detail varies across lexical items otherwise assumed to be phonetically identical. This detail is clearly maintainable: in changes in progress, items in the vanguard remain in the vanguard in the next generation, and highly reduced variants of frequent expressions remain reduced, i.e. change is not undone or regularized away. Indeed, if phonetically advanced or otherwise nonconformist expressions were subject to such policing, incipient changes, including those leading to grammaticalization, would always be thwarted, and languages and dialects would be static over time.

[6] To our knowledge, only Lüdtke (1980a,b;1985) has attempted to articulate a comprehensive theory of change incorporating the primary observation concerning reduction in internal change.

[7] It is important to emphasize that we are here referring to 'purely' phonetic augmentation of original lexical material. Morphological augmentation by means of affixation and compounding (which Matisoff 1990 views as providing 'compensatory phonological bulk' and Hopper 1994 as leading to 'compensatory accretion') is not phonetic or phonological (though, given sufficient elapsed time and dearth of historical information, it may appear to be).

[8] Note that what we are tracking in change is the history of particular items (the comparative method depends upon and requires cognates – specific form-meaning items, and not lexical items that are semantically or phonetically similar nor replacement items).

[9] Despite the apparent liberation from discrete bundles of features afforded by autosegmental approaches, the generative phonological enterprise remains fundamentally featural and segmental, and is thus true to, and bound by, its theoretical legacy (cf. Sampson 1974:237).

[10] Schemata of this sort are meant to represent the muscular event patterns generated by an idealized or 'average' vocal tract. This abstraction is not introduced to shield the theory from falsification, but rather from rejection on the basis of statistically insignificant deviation from the norm.

[11] If further justification for our choice of levels is needed, compare Harris (1974): "In spite of these complications, there seem to be three rather compelling reasons for studying the motor patterns of articulation. First, as Cooper 1965 and Liberman, Cooper, Shankweiler and Studdert-Kennedy have pointed out, they are one step closer to the speech-generating center in the brain than articulatory shapes. Second, most articulators are rather inaccessible – in some cases it may be easier to examine EMG signals than other physiological variables. Third, even when only anatomical data about the muscle are available, they allow some insight into the articulatory shapes which the speaker can generate" (p. 2283).

[12] In this way we might also account for the rapid extension of a diffusing sound change to non-core lexical items. Wang and Cheng (1977) note that changes appear to diffuse through the relevant lexicon in a gradual and uniform manner only in their early history, and then, after somewhat less than half of the candidate items have been affected, abruptly extend to the majority of remaining forms. If frequently-used words are typically in the vanguard in internal changes, it is possible that this phenomenon is at least in part interpretable in the following way. Changes proceed gradually through that portion of the lexicon roughly identifiable as 'core'. Core material serves as the basic input to acquisition, and hence also as that from which de-contextualized segmental (and longer) stretches are derived. Once a change has extended to all core expressions, the motor patterns for the old, pre-change state will no longer be available to new learners, who will derive, from the new, changed patterns, a new template for the production of peripheral material they later

encounter. Our view of substructures as derived over instances of larger holistic patterns seems to have some points of similarity with the exemplar model of categorization (Estes 1994).

[13] Measurements in Crystal & House (1988) show the mean duration of such flaps to be somewhat less than half that of 'hold-only' or unreleased [t]'s, more than 50% of a hold-only [d], and much shorter than complete stops.

[14] Of course, some might argue that this division is linguistically significant and thus hardly arbitrary.

[15] Cf. Browman & Goldstein's (1986) suggestion that gestural structures replace phonetic transcriptions.

[16] Various ways of dealing with the limitations of the segmental view have been explored by both phoneticians and phonologists. A particularly thorough discussion is given by Fowler (1980), who reviews earlier attempts and states the fundamental problem in the following terms. "What is not represented in the articulatory and acoustic records of an utterance is temporal discreteness. The different kinds of gestures go on simultaneously, and thus there are no borders perpendicular to the time axis in an articulatory or acoustic record to separate one segment from another" (p. 114). Cf. also Studdert-Kennedy (1987): "Since the earliest spectrographic, cineradiographic, and electromyographic studies, we have known that neither the articulatory nor the acoustic flow of speech can be divided into a sequence of segments corresponding to the invariant segments of linguistic description" (p. 68). Both Fowler's and Studdert-Kennedy's attempts at resolving the problem, however, are substantially different from our own.

[17] It is important to note here that we are not rejecting the idea of supralexical structure altogether. Rather, we are suggesting that it derives from, rather than determines, substance (the phonetic form of linguistic material) – that, in general terms, substance changes, and epiphenomenally, and in its wake, structure changes. As we have indicated above, the phonotactic patterns of a given dialect, for example, should be straightforwardly derivable as generalizations over the motor patterns of individual utterances in that dialect; these derived patterns may then function as motor templates. This view of the relation between substance and structure is compatible, in spirit and in some particulars, with Hopper's ideas on 'emergent grammar' (Hopper 1987).

[18] As Thomason, who has striven to develop a set of principles for distinguishing internal from external change generally, points out, the problem of determining internal vs. external origin is especially acute in cases of language shift in which the original language of the adopting community falls entirely into disuse (1986:244; see also Thomason & Kaufman 1988).

[19] A possible case in English dialects is the occurrence of an epenthetic vowel between the interdental and the liquid in words like *athlete* and *arthritis*. But not all rare or absent clusters need be so treated, as is illustrated by the apparent tolerance in English of *sphere* and *sphinx*, which involve independent lingual and labial articulations rather than the sequencing of different lingual configurations.

[20] Since vowels preceding nasal consonants are always at least partially nasalized, we are here referring to increase in nasalization on the vowel beyond the level of coarticulatory nasalization found synchronically. Coarticulatory patterns here, and hence the extent to which such vowels are nasalized, vary by language and dialect, and thus also evolve; our concern here is with the diachronic relation between nasality over the vowel and the fate of the original nasal consonant.

[21] Bhat (1974:41ff.) provides examples of palatalization entailing absorption in both IE and non-IE languages. Other cases are periodically encountered in grammatical

descriptions. For example, the progressive palatalization of /n/ following syllabic and non-syllabic /i/ in the Basque dialect of Getxo also entails complete absorption of the palatal vowel or glide (Hualde & Bilbao 1992:7).

[22] Another source is clearly methodological. The voicing contrasts for a proto-language are typically reconstructed *before* tonal patterns are considered. Having settled on the voicing contrasts the analyst then proceeds to determine if consonants influence tone or vice versa. But if the consonant types are determined prior to the determination of the possibility of tonal influence, Schuh's (1978) statement that "virtually no clear cases of tonal influence on segments have been found" is hardly surprising.

[23] See also Lass (1978). Our inclusion of 'Qualities' in Fig. 3 is intended to cover not only those consonantal residues which may temporally reduce onto surrounding vowels, but also those cases in which temporal and substantive reduction yield compensatory lengthening. We suspect that vowels are not actually *lengthened* in processes transcriptionally rendered as VC > V: . Rather, in those cases where the consonantal residues are insufficient to move supraglottal articulators (i.e. where only or primarily laryngeal residues remain) the supraglottal tract will be governed solely by inertial forces generated by temporally prior muscular events and by tissue elasticity.

[24] What we are doing here will seem at once overly complex and low level to traditional phonologists and overly simplified and facile to phoneticians who have had to deal with the problems of determining speech muscle activity by electromyographic means. Nevertheless, some attempt at bridging the gulf must be made. Further, we see no reason why phonologists armed with a decent anatomical handbook should not formulate laboratory-testable hypotheses about the nature and change of articulatory patterns. Whether or not these hypotheses are supported in the laboratory, the general field of inquiry should benefit from the procedure.

[25] To date there have been few electromyographic studies of the styloglossus. Smith (1971) examined styloglossus activity for velar consonants and for vowels, but not for the more anterior consonants. Smith's data show decreasing styloglossus activity for palatal [k'] versus velar [k]. This is consistent with the findings of Baer et al. (1988) for vowels, which show decreasing activity along the posterior-to-anterior axis. These EMG results are explicable solely on anatomical grounds (see, e.g. Miyawaki 1974). The hypothesis that styloglossus activity also plays a role in the articulation of alveolars and dentals is similarly warranted on anatomical grounds, but not yet supported by laboratory experimentation.

[26] The coordinative structures approach (Kelso et al. 1984, 1985, and other works) may reveal dependency relationships of this sort. This framework, however, is formulated with different objectives than our own: they wish to understand "both the *stability* of the spatiotemporal orchestration among gestures and the dynamical control structures that underlie such patterns" (Kelso et al. 1985:173; emphasis added RM/WP). Note that this is a strictly synchronic enterprise. Our goal, on the other hand, is to offer a constrained theory of how the muscular activity which underlies coordinative structures can change over time. The two goals are not incompatible, but their interface is not presently obvious. It may well be that individual muscular events will be found to group together in a manner consistent with the coordinative structures of Kelso et al., but until such groupings can be shown to be relevant to diachronic concerns, we take a more conservative stance.

[27] Diachronic equilibrium may be reflected synchronically as a "principle of energy consumption per unit time in muscle activity" (McAllister, Lubker & Carlson 1974:277) or as a "power constraint limiting energy expenditure per unit time" (Lindblom 1983:231). Note that an apparent alternative interpretation of (15b) – that

(the effects of) reductive events are precisely offset by the equivalent in augmentative events – accords poorly with the evidence of the historical record. For those languages for which we have attestation over some centuries, the long-term history of individual lexical items is characterized not by the maintenance but by the diminution of their energy and length, and by their gradual replacement by derivational forms and compounds (in which the original item may form the core) and periphrastic expressions, all of which in turn undergo reduction.

[28] See also Bloomfield (1933), Ohala (1974b), Anderson (1976), Coates (1980).

[29] Also *ny, *ng > ŋ > g.

[30] Dobson (1968) cites n > nd word-finally, at least for *sound;* Matthews (1938) cites n > nd, n > nt in some words in Cockney – 'surgent', 'orfunt', 'kindsman', 'sermonts'; cf. also *riband* as a variant of *ribbon*. Lewis & Pedersen (1974) note m > mp in the first plural endings of verbs and prepositions in Breton.

[31] Assuming that what we're seeing here are not in fact remnants of an earlier *d.

[32] *Hoffentlich* is also exceptional in not according with the voicing prediction of the first condition, but Clements presumably regards the violation here as the non-problematic result of German syllable-final obstruent devoicing.

[33] We may here briefly address the type B examples Clements cites as lacking phonetic motivation: lr > ldr (Sp. *saldrá*, FUT of *salir* 'leave'); nl > ŋgl (Fr. *épingle* 'pin' (<* *épindle*); and sl > skl, for which he cites no examples. For two of these, articulatory bases are by no means unimaginable. Given sufficient central contact in [r], its temporal reduction with [l] may well result in an [ldr] sequence, much as a stop arises from the coarticulation of central and lateral closure in ls sequences. The strangeness of nl > ŋgl may also be only apparent: if [n] first decayed to [ŋ] (as it does syllable-finally in many Spanish dialects of the Americas, Canfield 1981:7f.) this would result in ŋl, which could then develop directly to ŋgl. This would leave sl > skl as the only truly problematic case. The standard example seems to be Byzantine Greek *sklavinos* > *sklabos* ('Slav, Slavic' (from an original Slavic form with initial sl- with the same meaning), later 'slave'), and hence Medieval Latin *sclavus*. As a Greek rendering of an original Slavic word, the form with [k] thus cannot be regarded as the result of an evolutionary change. (Phelps 1937, on the basis of the north-south distribution of cognates with *sl and stl in Indo-European, argues for language contact for stl, as well as for later forms with skl and spl.).

[34] For an EMG study of Swedish inrounding, see McAllister, Lubker & Carlson (1974).

[35] Malmberg (1959) cites additional characteristics of Swedish vowels conducive to the analysis we describe here. The importance of this paper to our viewpoint lies in the detailed phonetic descriptions it provides. Were such detail regularly available to phonologists doing reconstruction, we believe that epenthetic solutions would be offered far less often. Unfortunately, all too often orthography is our only resource. This orthography must be interpreted, and our reconstructions guided by what we can observe on a shallow diachronic level and what we can infer on the basis of the statistical preponderance of reductive changes cross-linguistically.

[36] As a final illustration of this, consider Clements' (1992) challenge to Browman & Goldstein's 'temporal overlap and reduction only' hypothesis, in which he cites two particulars of NC sequences in Shona which seem not to be accounted for on a strictly phonetic view of assimilation (p. 189f.). He reproduces palatograms from Doke (1931) showing that nasals before dental and alveolar fricatives have full stop closure, and notes that Doke explicitly characterizes the assimilated nasal before [v] as a true bilabial. Arguing that temporal overlap cannot account for either full nasal stop closure before fricatives or for a true bilabial nasal before [v], Clements concludes that

"nasals assimilate constriction location, but not constriction degree, from a following continuant" (p. 190). In response, Browman & Goldstein point out that in order to adequately evaluate the situation, more detailed articulatory information on the forms in question is necessary. They nevertheless suggest that, to accommodate the evidence for oral closure in nasals before fricatives, articulatory phonology might be modified by specifically 'licensing' nasal-fricative sequences in which constriction degree differs. We take a stronger stand, and suggest that if these are evolutionary changes, they must have arisen from substantive and temporal reduction. Even without access to the history of Shona, it is clear that the current situation could have come about in a number of ways consistent with a phonetic approach. Though the bilabial character of the nasal before [v] is indisputable, Doke notes that [v] is rare in Shona, but [b] is common (1967:32). Rarity may indicate, among other things, recent innovation or occurrence restricted to loanwords. If loans are involved, we may be seeing a reflection of their ongoing assimilation to Shona patterns; that the labiodental nasal which occurs in other dialects is absent in Shona (Doke 1967:32) increases the plausibility of such a scenario. If [v] is rather fully native, then, since it is certainly not impossible that it is historically related to [b], the bilabiality of the nasal may be a retention from the assimilation induced by original stem-initial bilabial fricatives, which subsequently reduced to labiodentals; if [v] in fact arose subsequent to the assimilation to [m], there is no anomaly. A reduction of [mb] to [mv] would also be consistent with the occurrence in Shona of [pf] and [bv] (Doke (p. 33) cites Shona as one of three dialects in which labial affricates are not fully homorganic). As for [n] before [s], despite Clements' remark that "[...] there is no evidence that the nasal component is associated with a tongue tip gesture of its own at any level of description; in phonological terms, it constitutes a prenasalization element" (p. 189), it must be remembered that while an unspecified nasal may be fine as a synchronic descriptive tool, historically there had to have been a fully articulated nasal here, and if it was originally [n] (before assimilation), then closure is to be expected in its descendant. In this case, there is no increase in closure before [s], and consequently no event at all, but rather simple retention of the original nasal.

[37] The understandable desire for elegance or economy in description may, via reliance on distributional patterns, also be the source of certain cases of apparent metathesis, especially as posited for languages and families for which no historical information is available. Consider a scenario in which a parent dialect (now extinct) had $-C_1VC_2-$ sequences, where $C_1 = C_2$. If one surviving daughter dialect had reduced C_1 to zero and another had reduced C_2 to zero, then the resulting patterns (VC and CV respectively) would, in the absence of knowledge of the parental situation, just as easily (and perhaps preferably) be regarded as having come about via a single event – metathesis – in one daughter, with the other daughter preserving a presumed parental CV or VC.

[38] The extent to which the authority of a prestige dialect is taken for granted may be made clear by the following example. The comparativist Louis Gray (1934:25) begins his discussion of vocalic epenthesis in Semitic by citing, as one of the non-Semitic examples offered as an illustration, 'vulgar' English *chiminey* (cf. Std. Eng. *chimney*). The internasal vowel, however, is original (Lat. *camīnus* 'forge, hearth' > Late L. (*camera*) *camina–ta* '(a room) provided with a chimney' > OF *cheminee* > ME *chimenee*); the non-standard form may then indicate retention. As we noted above, unstressed medial vowels such as this have been reducing to zero in an item-specific manner for some centuries (cf. Jespersen 1909; also Zwicky 1972, Hooper 1976b; and Kökeritz 1969 on the reinstatement of similar weak vowels as a result of spelling pronunciation).

[39] Here, too, we may question whether an epenthetic analysis would have been proposed had fate treated the ancestors of Maru- and Letsi-speaking peoples more

kindly, i.e. if today there were greater numbers of Tibeto-Burman daughter dialects with consonantal finals, spoken by large numbers of people.

[40] Clear cases of this are given by Shafer (1974:345f.) for Burmese dialects. Shafer devotes a full page of text and accompanying tables to clearly document numerous instances where fieldworkers transcribed final glottal stops as [t]'s and [k]'s and final nasalized vowels as vowel + nasal consonant sequences.

[41] Haiman (1994) explores some of the parallels between grammaticalization and the ritualization of behavior in other species.

References

ALLPORT A., D. MACKAY, W. PRINZ, & E. SCHEERER, eds. (1987), *Language perception and production*, London, Academic Press.

ANDERSEN H. (1972), "Diphthongization", *Language* 48:11-50.

ANDERSEN H. (1973), "Abductive and deductive change", *Language* 49:765-793.

ANDERSON J.M. & C. JONES (1974), *Historical linguistics II: Theory and Description in phonology*, Amsterdam, North-Holland.

ANDERSON S.A. (1976), "Nasal consonants and the internal structure of segments", *Language* 52:326-344.

ANTTILA R. (1972), *An Introduction to Historical and Comparative Linguistics*, New York, Macmillan.

BAER T., P. ALFONSO & K. HONDA (1988), "Electromyography of the tongue muscles during vowels in /əpVp/ environment", *Research Institute of Logopedics and Phoniatrics, School of Medicine, University of Tokyo, Annual Bulletin* 22:7-19.

BAILEY C.-J.N. (1973), *Variation and Linguistic Theory*, Arlington, VA, Center for Applied Linguistics.

BAILEY C.-J.N. (1977), "Linguistic change, naturalness, mixture, and structural principles", *Papiere zur Linguistik* 16:6-73.

BARRACK C.M. (1976), "Lexical diffusion and the High German consonant shift", *Lingua* 40:151-175.

BEE D. (1973), "Comparative and historical problems in East New Guinea Highland languages", in MCKAUGHAN (1973:694-738).

BENEDICT P. (1972), *Sino-Tibetan: a Conspectus*, Cambridge, University Press.

BHAT D.N.S. (1974), "A general study of palatalization", *WPLU* 14:17-58.

BLOOMFIELD L. (1933), *Language*, New York, Holt, Rinehart & Winston.

BOAS F. (1911), Introduction to the *Handbook of American Indian languages*, *BAE Bulletin* 40, Part 1. Reprinted 1963, Washington D.C., Georgetown University Press.

BRASINGTON R.W.P. (1973), "Reciprocal rules in Catalan phonology", *Journal of Linguistics* 9:25-33.

BROWMAN C.P. & L. GOLDSTEIN (1986), "Towards an articulatory phonology", *Phonology Yearbook* 3:219-252.

BROWMAN C.P. & L. GOLDSTEIN (1990), "Tiers in articulatory phonology, with some implications for casual speech", in KINGSTON & BECKMAN (1990:341-376).

BROWMAN C.P. & L. GOLDSTEIN (1991), "Gestural structures: distinctiveness,

phonological processes, and historical change", in MATTINGLY & STUDDERT-KENNEDY (1991:313-338).

BROWMAN C.P. & L. GOLDSTEIN (1992), "Articulatory phonology: an overview", *Phonetica* 49:155-180.

BUCK C.D. (1904), *A Grammar of Oscan and Umbrian, with a collection of inscriptions and a glossary*, Boston, Ginn and Company.

BURLING R. (1966), "The addition of final stops in the history of Maru (Tibeto-Burman)", *Language* 42:581-586.

BURLING R. (1967), "Proto Lolo-Burmese", *IJAL* 33, No. 2, Part 2.

BUTTERWORTH B., ed., (1980), *Language production, Vol 1: Speech and Talk*, London, Academic Press.

BUTTERWORTH B., B. COMRIE & Ö. DAHL, eds., (1984) *Explanations for language universals*, Berlin, Mouton.

BYBEE J., W. PAGLIUCA & R. PERKINS (1991), "Back to the future", in TRAUGOTT & HEINE (1991, 2:17-58).

BYBEE J., R. PERKINS & W. PAGLIUCA (1994), *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*, Chicago, University of Chicago Press.

CANFIELD D.L. (1981), *Spanish Pronunciation in the Americas*, Chicago, University of Chicago Press.

CATFORD J.C. (1977), *Fundamental Problems in Phonetics*, Bloomington, IN, Indiana University Press.

CATFORD J.C. (1991), "Contribution to panel discussion", in MATTINGLY & STUDDERT-KENNEDY (1991:176-177).

CHEN M. & H. CLUMECK (1975), "Denasalization in Korean: a search for universals", in FERGUSON, HYMAN & OHALA (1975:125-131).

CHEN M. & W.S-Y. WANG (1975), "Sound change: actuation and implementation", *Language* 51:255-281.

CHOMSKY N. & M. HALLE (1968), *The Sound Pattern of English*, New York, Harper & Row.

CHRISTIE W., ed. (1976), *Current Progress in Historical Linguistics*, Amsterdam, North-Holland.

CLEMENTS G.N. (1987), "Phonological feature representation and the description of intrusive stops", *CLS* 23/2:29-50.

CLEMENTS G.N. (1992), "Phonological primes: features or gestures?", *Phonetica* 49:181-193.

COATES R. (1980), "Time in phonological representations", *Journal of Phonetics* 8:1-20.

COOPER F. (1965), "Research techniques and instrumentation: EMG", *Proceedings of the conference Communicative Problems in Cleft Palate. American Speech and Hearing Association Report* 1:153-168.

COOPER W.E. & J. PACCIA-COOPER (1980), *Syntax and Speech*, Cambridge, MA, Harvard University Press.

CRYSTAL T.H. & A.S. HOUSE (1988), "Segmental durations in connected-speech signals: current results", *JASA* 83:1553-1573.

DOBSON E.J. (1968), *English Pronunciation 1500-1700*, 2 vol., 2nd ed. Oxford, Clarendon Press.

DOKE C.M. (1931), *A Comparative Study of Shona Phonetics*, Johannesburg, University of the Witwatersrand Press.

DOKE C.M. (1967), *The Southern Bantu Languages*, London, International African Institute.
ESTES W.K. (1994), *Classification and Cognition*, New York/Oxford, Oxford University Press.
FERGUSON C.A. & C.B. FARWELL (1975), "Words and sounds in early language acquisition: English initial consonants in the first fifty words", *Language* 51:419-439.
FERGUSON C.A., L. HYMAN & J. OHALA, eds. (1975), *Nasálfest: Papers from a symposium on nasals and nasalization*, Stanford, Stanford University.
FIDELHOLZ J. (1975), "Word frequency and vowel reduction in English", *CLS* 11:200- 213.
FISCHER-JØRGENSEN E. (1968), "Les occlusives francaises et danois d'un sujet bilingue", *Word* 24:112-153.
FISIAK J., ed. (1978), *Recent Developments in Historical Phonology*, The Hague, Mouton.
FISIAK J., (1985), *Historical Semantics, Historical Word Formation*, Berlin, Mouton de Gruyter.
FOLEY J. (1970), "Phonological distinctive features", *Folia Linguistica* 4:87-92.
FOURAKIS M. & R. PORT (1986), "Stop epenthesis in English", *Journal of Phonetics* 14:197-221.
FOWLER C.A. (1980), "Coarticulation and theories of extrinsic timing", *Journal of Phonetics* 8:113-33.
FOWLER C.A., P. RUBIN, R.E. REMEZ & M. TURVEY (1980), "Implications for speech production of a general theory of action", in BUTTERWORTH (1980:373-420).
FROMKIN V., ed. (1978), *Tone: a Linguistic Survey*, New York, Academic Press.
FROMKIN V. (1985), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, Orlando, FL, Academic Press.
GAIRDNER W.H.T. (1925), *The Phonetics of Arabic*, London, Oxford University Press.
GIACALONE RAMAT A., O. CARRUBA & G. BERNINI, eds. (1987), *Papers from the 7th International Conference on Historical Linguistics*, Amsterdam, John Benjamins.
GRANDGENT C.H. (1927), *From Latin to Italian: an Historical Outline of the Phonology and Morphology of the Italian Language*, Cambridge, MA, Harvard University Press.
GRAY L.H. (1902), *Indo-Iranian Phonology, with Special Reference to the Middle and New Indo-Iranian Languages*, New York, Columbia University Press.
GRAY L.H. (1934), *Introduction to Semitic Comparative Linguistics*, New York, Columbia University Press.
HAAS M. (1969),"Internal reconstruction of the Nootka-Nitinat pronominal suffixes", *IJAL* 35:108-124.
HAIMAN J. (1994), "Ritualization and the development of language", in PAGLIUCA (1994: 3-28).
HALLE M. (1983), "On distinctive features and their articulatory implementation", *NLLT* 1:91-105.
HALLE M. (1985), "Speculations about the representation of words in memory", in Fromkin (1985:101-114).
HALLE M. & G.N. CLEMENTS (1983), *Problem Book in Phonology*, Cambridge, MA, MIT Press.

HALLE M. & K. STEVENS (1971), "A note on laryngeal features", *Quarterly Progress Report, Research Laboratory of Electronics, MIT* 101:198-213.
HAN M.S. (1969), "Vietnamese tones", *Studies in the phonology of Asian languages VIII*. Los Angeles, Acoustic Phonetics Research Laboratory, University of Southern California.
HARRIS K.S. (1973), "Stress and syllable duration change", *Haskins Laboratories SR-* 35/36:31-38.
HARRIS K.S. (1974), "Physiological aspects of articulatory behavior", in SEBEOK (1974:2281- 2302).
HOCKETT C.F. (1955), *A Manual of Phonology*, *IJAL* Memoir 11.
HOCKETT C.F. (1965), "Sound change", *Language* 41:185-204.
HOIJER H. (1966), "Hare phonology: an historical study", *Language* 42:499-507.
HOMBERT J-M., ed. (1976), *Studies on Production and Perception of Tones*, *UCLA WPP* 33.
HOMBERT J-M., J.J. OHALA & W. EWAN (1979), "Phonetic explanations for the development of tones", *Language* 55:37-58.
HOOPER J.B. (1976a), An *Introduction to Natural Generative Phonology*, New York, Academic Press.
HOOPER J.B. (1976b), "Word frequency in lexical diffusion and the source of morpho- phonological change", in CHRISTIE (1976:95-105).
HOOPER J.B. (1981), "The empirical determination of phonological representations", in MYERS, LAVER & ANDERSON (1981:347-357).
HOPPER P. (1987), "Emergent grammar", *BLS* 13:139-157.
HOPPER P. (1994), "Phonogenesis", in PAGLIUCA (1994:29-45).
HUALDE J.-I. & X. BILBAO (1992), *A phonological study of the Basque dialect of Getxo*, *Annuario del Seminario de Filología «Julio de Urquijo»*, Supplement XXIX.
HYMAN L.M. (1973), "The feature [grave] in phonological theory", *Journal of Phonetics* 1:329-37.
HYMAN L.M. (1976), "On some controversial questions in the study of consonant types and tone", in HOMBERT (1976:90-98).
HYMAN L.M. & R. SCHUH (1974), "Universals of tone rules: evidence from West Africa", *Linguistic Inquiry* 5:81-115.
JAVKIN H. (1978), "Phonetic and grammatical explanations for an epenthesis and a non-epenthesis in English", *BLS* 4:330-341.
JEANJAQUET J. (1931), "Les patois valaisans: caractères généraux et particularités", *Revue de linguistique Romane* 7:23-51.
JESPERSEN O. (1909), *A Modern English Grammar on Historical Principles. Part I: Sounds and spelling*, Copenhagen/London, Einar Munksgaard and George Allen & Unwin.
JESPERSEN O. (1922), *Language: its Nature, Development, and Origin*, London, George Allen & Unwin.
KAPLAN H.M. (1971), *Anatomy and Physiology of Speech*, 2nd ed. New York, McGraw-Hill.
KEATING P. (1985a), "CV phonology, experimental phonetics, and coarticulation", *UCLA WPP* 62:1-13.
KEATING P. (1985b), "The phonetics-phonology interface", *UCLA WPP* 62:14-33.
KELLER E. (1987), "The cortical representation of motor processes of speech", in KELLER & GOPNIK (1987:125-162).

KELLER E. & M. GOPNIK, eds. (1987), *Motor and Sensory Processes of Language*, Hillsdale, NJ & London, Lawrence Erlbaum Associates.
KELSO J.A.S., E. SALZMAN & B. TULLER (1985), "The dynamical perspective on speech production: data and theory", *Haskins Laboratories SR*-84:171-205.
KELSO J.A.S., B. TULLER, E. VATIKIOTIS-BATESON & C.A. FOWLER (1984), "Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures", *Journal of Experimental Psychology: Human Perception and Performance* 10:812-832.
KINGSTON J. & M.E. BECKMAN, eds. (1990), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, Cambridge, University Press.
KIPARSKY P. (1965), *Phonological Change*, Dissertation, MIT, IULC 1971.
KÖKERITZ H. (1969), "Elizabethian prosody and historical phonology", in LASS (1969: 208-227).
KRISHNAMURTI BH. (1978), "Areal and lexical diffusion of sound change: evidence from Dravidian", *Language* 54:1-20.
LABOV W. (1974), "On the use of the present to explain the past", *Proceedings of the 11th International Congress of Linguists*, 825-851. Bologna, Società Editrice Il Mulino.
LABOV W. (1981), "Resolving the Neogrammarian controversy", *Language* 57:267-308
LABOV W., M. YAEGER & R. STEINER (1972), *A Quantitative Study of Sound Change in Progress*, Philadelphia, US Regional Survey.
LADEFOGED P. (1971), *Preliminaries to Linguistic Phonetics*, Chicago, University of Chicago Press.
LADEFOGED P. (1980), "What are linguistic sounds made of?", *Language* 56:485-502.
LASS R. (1976), *English Phonology and Phonological Theory*, Cambridge, University Press.
LASS R. (1978), "Mapping constraints in phonological reconstruction: on climbing down trees without falling out of them", in FISIAK (1978:245-286).
LASS R. (1984), *Phonology*, Cambridge, University Press.
LASS R., ed. (1969), *Approaches to English Historical Linguistics*, New York, Holt, Rinehart & Winston.
LEHMANN W.P. & Y. MALKIEL, eds. (1968), *Directions for Historical Linguistics*, Austin, University of Texas Press.
LEWIS H. & H. PEDERSEN (1974), *A Concise Comparative Celtic Grammar*, 3rd ed. Göttingen, Vandenhoeck & Ruprecht.
LIBERMAN A.M., F.S. COOPER, D.P. SHANKWEILER & M. STUDDERT-KENNEDY (1967), "Perception of the speech code", *Psychological Review* 74:431-461.
LIBERMAN I.Y., D. SHANKWEILER, F.W. FISCHER & B. CARTER (1974), "Explicit syllable and phoneme segmentation in the young child", *Journal of Experimental Child Psychology* 18:201-212.
LIEBER R. (1987), *An Integrated Theory of Autosegmental Processes*, Albany, SUNY Press.
LINDBLOM B. (1983), "Economy of speech gestures", in MACNEILAGE (1983:217-45).
LINDBLOM B., P. MACNEILAGE & M. STUDDERT-KENNEDY (1984), "Self-organizing

processes and the explanation of phonological universals", in BUTTERWORTH, COMRIE & DAHL (1984: 181-203).
LÜDTKE H. (1980a), "Sprachwandel als universales Phänomen", in LÜDTKE (1980:1-19).
LÜDTKE H. (1980b), "Auf dem Weg zu einer Theorie der Sprachwandels", in LÜDTKE (1980: 182-252).
LÜDTKE H., ed. (1980), *Kommunikationstheoretische Grundlagen der Sprachwandels*, Berlin, Walter de Gruyter.
LÜDTKE H. (1985), "Diachronic irreversibility in word-formation and semantics", in FISIAK (1985:355-366).
MACNEILAGE P.F. (1970), "Motor control of serial ordering of speech", *Psychological Review* 77:182-196.
MACNEILAGE P.F., ed. (1983), *The Production of Speech*, New York, Springer.
MADDIESON I. (1974), "A note on tone and consonants", in MADDIESON, ed. (1974:18-27).
MADDIESON I., ed. (1974), *The Tone Tome: Studies on Tone from the UCLA Tone Project, UCLA WPP* 27.
MALMBERG B. (1959), "Les voyelles suedoises et la notion de 'fermeture vocalique'", *Studia Linguistica* 13:49-61.
MALMBERG B. (1963), *Phonetics*, New York, Dover.
MALONE J.L. (1971), "Systematic metathesis in Mandaic", *Language* 47:394-415.
MARAN L.R. (1971a), "A note on the development of tonal systems in Tibeto-Burman", *Occasional Papers of the Wolfenden Society on Tibeto-Burman Linguistics* 2:1-24.
MARAN L.R. (1971b), "Burmese and Jingpho: a study of tonal linguistic processes", *Occasional Papers of the Wolfenden Society on Tibeto-Burman Linguistics* 4.
MATISOFF J.A. (1968), "Review of Burling 1967", *Language* 44:879-897.
MATISOFF J.A. (1990), "Bulging monosyllables: areal tendencies in Southeast Asian diachrony", *BLS* 15:543-559.
MATTHEWS W. (1938), *Cockney Past and Present: a Short History of the Dialect of London*, New York, Dutton.
MATTINGLY I.G. & M. STUDDERT-KENNEDY, eds. (1991), *Modularity and the Motor Theory of Speech Perception: Proceedings of a Conference to Honor Alvin M. Liberman*, Hillsdale, NJ, Lawrence Erlbaum.
MCALLISTER R., J. LUBKER & J. CARLSON (1974), "An EMG study of some characteristics of the Swedish rounded vowels", *Journal of Phonetics* 2:267-278.
MCKAUGHAN H., ed. (1973), *The Languages of the Eastern family of the East New Guinea Highland Stock*, Seattle, University of Washington Press.
MENÉNDEZ PIDAL R. (1968), *Manual de gramática histórica española*, Décimotercera edición, Madrid, Espasa-Calpe.
MILLARDET G. (1911), "Insertions de consonnes en suedois moderne", *Revue de phonetique* 1:309-346.
MILLER R.A. (1968), "Once again, the Maru final stops", Paper read at the First conference on Sino-Tibetan, Yale University, October.
MIYAWAKI K. (1974), "A study of the musculature of the human tongue (observations on transparent preparations of serial sections)", *Research Institute of Logopedics and Phoniatrics, School of Medicine, University of Tokyo, Annual Bulletin* 8:23- 50.

MORAIS J. (1985), "Literacy and awareness of the units of speech: implications for research on the units of perception", *Linguistics* 23:707-721.
MORAIS J., L. CARY, J. ALEGRIA & P. BERTELSON (1979), "Does awareness of speech as a sequence of phones arise spontaneously?", *Cognition* 7:323-331.
MOWREY R. & I.R.A. MACKAY (1990), "Phonological primitives: electromyographic speech error evidence", *JASA* 88:1299-1312.
MOWREY R. & W. PAGLIUCA (1987), "Epenthesis in evolutionary and non-evolutionary change", Paper presented at the 8th International Conference on Historical Linguistics, Lille, France.
MOYNIHAN M. (1970), "Control, suppression, decay, disappearance, and replacement of displays", *Journal of Theoretical Biology* 29:85-112.
MYERS T., J. LAVER & J. ANDERSON, eds. (1981), *The cognitive Representation of Speech*, Amsterdam, North-Holland.
NAVARRO T. (1968), *Studies in Spanish Phonology*, tr. by R. Abraham, Coral Gables, FL, University of Miami Press.
OHALA J.J. (1971), "Physical models in phonology", *Proceedings of the 7th International Congress of Phonetic Sciences*, Montreal, 1166-1171.
OHALA J.J. (1974a), "Experimental historical phonology", in ANDERSON & JONES (1974:353- 389).
OHALA J.J. (1974b), "Phonetic explanation in phonology", *CLS Parasession on Natural Phonology*, 251-274.
OHALA J.J. (1975), "Phonetic explanations for nasal sound patterns", in FERGUSON, HYMAN & OHALA (1975:289- 316).
OHALA J.J. (1978), "Southern Bantu vs. the world: the case of palatalization of labials", *BLS* 4:370-386.
OHALA J.J. (1981), "The listener as a source of sound change", *CLS Parasession on language and behavior*:178-203.
OHALA J.J. (1983a), "The direction of sound change", *Abstracts of the 10th International Congress of Phonetic Sciences*, 253-8. Dordrecht, Holland, and Cinnaminson, NJ, Foris.
OHALA J.J. (1983b), "The phonological end justifies any means", *Proceedings of the 8th International Congress of Linguists*, 232-43.
OHALA J.J. (1990),"The phonetics and phonology of aspects of assimilation", in KINGSTON & BECKMAN (1990:258-275).
PAGLIUCA W. (1982), *Prolegomena to a Theory of Articulatory Evolution*, Dissertation, State University of New York at Buffalo. Ann Arbor, University Microfilms International.
PAGLIUCA W., ed. (1994), *Perspectives on Grammaticalization*, Amsterdam, John Benjamins.
PAGLIUCA W. & R. MOWREY (1980), "On certain evidence for the feature [grave]", *Language* 56:503- 514.
PAGLIUCA W. & R. MOWREY (1987), "Articulatory evolution", in GIACALONE RAMAT ET AL. (1987:459- 472).
PETERS A.M. (1983), *The Units of Language Acquisition*, Cambridge, University Press.
PHELPS J. (1937), "Indo-European initial sl", *Language* 13:279-284.
PHILLIPS B.S. (1980), "Old English *an* ~ *on*: a new appraisal", *Journal of English Linguistics* 14:20-23.

PHILLIPS B.S. (1984), "Word frequency and the actuation of sound change", *Language* 60:320-342.
PIGGOTT G.L. & R. SINGH (1985), "The phonology of epenthetic segments", *Canadian Journal of Linguistics* 30:415-451.
POPE M.K. (1934), *From Latin to Modern French, with especial consideration of Anglo-Norman*, Manchester, Manchester University Press.
POSTAL P. (1968), *Aspects of Phonological Theory*, New York, Harcourt, Brace.
QUESADA F. (1988), *Phonological Processes in Quechua and their implications for Phonological Theory*, Dissertation, State University of New York at Buffalo. Lima, CONCYTEC.
ROCA I. M. (1975), "Phonetics is not phonology", *Journal of Phonetics* 3:53-62.
ROHLFS G. (1966), *Grammatica storica della lingua italiana e dei suoi dialetti: Fonetica*, Turin, Giulio Einaudi.
RUDES B.A. (1976), *Historical Phonology and the Development of the Tuscarora Sound System*, Ann Arbor, University Microfilms.
SAMPSON G. (1974), "Is there a universal phonetic alphabet?", *Language* 50:236-259.
SCHUCHARDT H. (1885), *Über die Lautgesetze: gegen die Junggrammatiker*, Berlin, Robert Oppenheim. Reprinted, with an English translation, in *Schuchardt, the Neogrammarians, and the Transformational Theory of Phonological Change: Four Essays by Hugo Schuchardt, Theo Vennemann, & Terence H. Wilbur*, Frankfurt, Athenäum.
SCHUH R. (1978), "Tone rules", in FROMKIN (1978:221-256).
SEBEOK T.A., ed. (1974), *Current Trends in Linguistics, Vol 12*, The Hague, Mouton.
SHAFER R. (1947), "Penutian", *IJAL* 13:205-219.
SHAFER R. (1974), *Introduction to Sino-Tibetan*, Wiesbaden, Otto Harrassowitz.
SHEVELOV G.Y. (1965), *A Prehistory of Slavic: the Historical Phonology of Common Slavic*, New York, Columbia University Press.
SIEVERS E. (1901), *Grundzüge der Phonetik zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen*, Leipzig, Breitkopf & Hartel.
SIVERTSEN E. (1960), *Cockney Phonology*, Oslo, Oslo University Press.
SMITH T.S. (1971), *A Phonetic Study of the Function of the Extrinsic Tongue Muscles*, *UCLA WPP* 18.
STOCKWELL R. & R.K.S. MACAULAY, eds. (1972), *Linguistic Change and Generative Theory*, Bloomington, Indiana University Press.
STUDDERT-KENNEDY M. (1987), "The phoneme as a perceptuomotor structure", in ALLPORT, MACKAY, PRINZ & SCHEERER (1987:67-84).
TAPPOLET E. (1931), "Les données fondamentales des condition linguistiques du Valais (Suisse)", *Revue de linguistique Romane* 7:9-22.
THOMASON S.G. (1986), "On establishing external causes of language change", *ESCOL* 2:243-251.
THOMASON S.G. & T. KAUFMAN (1988), *Language Contact, Creolization, and Genetic Linguistics*, Berkeley, University of California Press.
TRAUGOTT E. & B. HEINE, eds. (1991), *Approaches to Grammaticalization*, 2 vol., Amsterdam, John Benjamins.
VAGO R. (1976), "More evidence for the feature [grave]", *Linguistic Inquiry* 7:671-674.

WANG W.S-Y. (1969), "Competing changes as a cause of residue", *Language* 45:9-25.
WANG W.S-Y., ed. (1977), *The Lexicon in Phonological Change*, The Hague, Mouton.
WANG W.S-Y. & C-C. CHENG (1977), "Implementation of phonological change: the Shuang- Feng Chinese case", in WANG (1977:148-158).
WARREN R.M. (1976), "Auditory perception and speech evolution", *Ann. N.Y. Acad. Sci.* 280:708-717.
WARREN R.M. (1982), *Auditory Perception: a New Synthesis*, New York, Pergamon.
WEINREICH U. (1953), *Languages in Contact: Findings and Problems*. 2nd printing 1963, The Hague, Mouton.
WEINREICH U., W. LABOV & M.I. HERZOG (1968), "Empirical foundations for a theory of linguistic change", in LEHMANN & MALKIEL (1968:95-188).
WETZELS W.L. (1985), "The historical phonology of intrusive stops", *Canadian Journal of Linguistics* 30:285-333.
YIP M. (1987), "English vowel epenthesis", *NLLT* 5:463-484.
ZIPF G.K. (1929), "Relative frequency as a determinant of phonetic change", *Harvard Studies in Classical Philology* XL:1-95.
ZIPF G.K. (1932), *Selected Studies of the Principle of Relative Frequency in Language*, Cambridge, MA, Harvard University Press.
ZSIGA E.C. (1992), "Acoustic evidence for gestural overlap in consonant sequences", *Haskins Laboratories SR*-111/112:43-62.
ZWICKY A.M. (1972), "Note on a phonological hierarchy in English", in STOCKWELL & MACAULAY (1972:275-301).