5 The Generative Approach to Phonology

Introduction

This chapter deals with an approach to phonology which represents an influential alternative to the phonemic view of the previous chapter. After a brief account of the origins of generative phonology (5.1) and of Chomsky and Halle's major work *The sound pattern of English* (5.2), the heart of the chapter is devoted to explaining and illustrating the basic notation and principles of generative phonology:

- rule notation (5.3)
- formalism and evaluation (5.4)
- abbreviatory devices (5.5)
- rule order (5.6).

The final part of the chapter treats critical issues that have arisen in the elaboration of generative phonology:

- functional considerations (5.7)
- the notions of naturalness and markedness (5.8)
- abstractness in phonological description (5.9).

5.1 The origins of generative phonology

The 1960s saw increasing discontent with orthodox phonemics in North America. A series of publications by Halle (1959, 1962, 1964), a vigorous attack by Chomsky on phonemics and structuralist linguistics in general (1964), a book by Postal (1968), and a large-scale treatment of English phonology jointly authored by Chomsky and Halle (1968) marked the emergence of generative phonology as a new theory and framework of description.

Halle had been involved in research and publication on phonological features or components (chapter 10 below) and went on to devote attention to the function of features within phonological systems. In assessing phonological description – and particularly in formulating phonological rules – Halle argued that plausible general rules were better expressed in terms of features. A phonological process whereby all plosives are voiced between vowels is a plausible rule: it is known to operate in some languages and it seems to reflect a probable pattern of voicing assimilation. It is a more likely rule than one which says, for example, that [p] is voiced only between [a] and [u], [t] is voiced only between [u] and [i], and [k] is voiced only between [e] and [o].

Most phoneticians and phonologists readily agree that there are 'normal' tendencies in speech and that certain processes seem more common or more plausible than others – although their universality should not be exaggerated (section 4.1 above). Halle's point, however, concerns description and explanation: when expressed in segments, plausible rules do not necessarily appear simpler. The two rules suggested above might appear as

(5.1.1) $\left.\begin{matrix}[p] \rightarrow [b] \\ [t] \rightarrow [d] \\ [k] \rightarrow [g]\end{matrix}\right\}$ between $\left\{\begin{matrix}[i] \\ [e] \\ [a] \\ [o] \\ [u]\end{matrix}\right\}$ and $\left\{\begin{matrix}[i] \\ [e] \\ [a] \\ [o] \\ [u]\end{matrix}\right\}$

(5.1.2) [p] → [b] between [a] and [u]
[t] → [d] between [u] and [i]
[k] → [g] between [e] and [o].

Of course the first rule can be expressed as a general statement, such as

any voiceless plosive is voiced between any two vowels.

In this wording, it is the use of features (voiceless, plosive, and so on) that captures the generality of the rule. If we adopt the same style with (5.1.2), our use of features now makes the rule much more cumbersome than (5.1.1):

a voiceless bilabial plosive is voiced between a low vowel and a high back vowel; a voiceless alveolar plosive is voiced between a high back vowel and a high front vowel; and a voiceless velar plosive is voiced between a mid front vowel and a mid back vowel.

This, according to Halle, is precisely what we want – the more plausible general rule looks simpler, the less plausible looks more complex. In other words, phonological description should employ feature-based rules as a proper means of reflecting the complexity of the description. This does not mean, of course, that rules such as (5.1.2) are said to be impossible, only that they are far less likely than rules such as (5.1.1) and that it is therefore proper to signal their complexity.

The use of rules and features as the elements of phonological description meant that the concept of the phoneme was under threat. Indeed, Halle claimed that the phoneme was often a hindrance to description. In his treatment of Russian phonology (1959), he cited an example which has been quoted in subsequent literature repeatedly (*ad nauseam*, according to Sommerstein 1977, p. 116). In brief, Halle points out that there is a general rule in Russian that an obstruent (plosive or fricative) is voiced when preceding a voiced obstruent. Thus a word-final voiceless plosive will be voiced if the following word begins with a voiced plosive: [t] + [b] is pronounced as [d] + [b], [p] + [g] as [b] + [g], and so on. Now, in orthodox phonemic terms Russian has distinct voiced and voiceless plosive phonemes. We find, for instance, /bil/ ('was') versus /pil/ ('blaze, glow'), /djenj/ ('day') versus /tjenj/ ('shade, shadow'), as minimal pairs. But Russian does not have voiced and voiceless affricates as separate phonemes: there is no phonemic contrast between [tʃ] and [ʤ] or between [ts] and [dz], and the voiced affricates are simply allophones of their voiceless counterparts. Hence, in a phonemic account, when a word-final /t/ is voiced preceding a voiced obstruent, we are dealing with the substitution of /d/ for /t/, of one phoneme for another. On the other hand, when a word-final /ts/ affricate is voiced in the same context, /ts/ is realized as its voiced allophone [dz]. But, Halle argues, the phenomenon of voicing assimilation in Russian is surely a single process, and not one of phonemic substitution in some cases and allophonic conditioning in others. We should be suspicious of a framework of description which leads us to an awkward account of such an apparently straightforward phenomenon. We ought to be able to say that Russian simply has a phonological rule that obstruents are voiced when preceding voiced obstruents.

Postal (1968, pp. 36–7) gives another example designed to undermine the centrality of the phoneme. In Mohawk, it can happen that /t/ or /k/ precedes /j/ across a morpheme boundary, but both sequences are realized as [ʤ]. Postal argues that it should be legitimate to say that [ʤ] is derived, by rule, from two different sources, namely /tj/ and /kj/. This of course makes [ʤ] phonologically ambiguous, in violation of the biuniqueness principle (section 4.9 above). And it is not clear how a phonemic account can satisfactorily avoid this violation. It would be possible to say that [ʤ] unambiguously represents /tj/ and that /kj/ becomes /tj/ by morphophonemic rule, but Postal points to the arbitrariness of this decision. Why doesn't [ʤ] realize /kj/, with /tj/ becoming /kj/ by morphophonemic rule? Postal's solution, in the spirit of generative phonology, is to dispense with the phonemic level and morphophonemic rules altogether. If we regard /tj/ and /kj/ as rather deeper or more abstract than a phonemic transcription, then we can state relatively neat and general phonological rules which derive the phonetic forms from these underlying representations.

Arguments of this kind led generative phonologists to abandon the concepts of phoneme and allophone, and to talk in terms of a relatively abstract or morphophonemic underlying level of phonological representation from which the phonetic output could be derived by application of a set of phonological rules. The elaboration of this new conception of phonology was part of the development of the transformational–generative theory of language in general, pioneered by Noam Chomsky. Although he is sometimes thought of as a grammarian

with a particular interest in syntax, Chomsky himself contributed to the development of generative phonology. His *Current issues in linguistic theory* (1964) is generally critical of modern linguistics: the nineteenth century narrowed 'the scope of linguistics to the study of inventory of elements' (p. 22), and de Saussure and 'structural linguistics' were preoccupied with 'systems of elements rather than the systems of rules which were the focus of attention in traditional grammar' (p. 23). Against this background he dismisses much of modern phonology as 'taxonomic phonemics', having referred to 'a curious and rather extreme contemporary view to the effect that true linguistic science must *necessarily* be a kind of pre-Darwinian taxonomy concerned solely with the collection and classification of countless specimens' (1964, p. 25). He criticizes in detail (pp. 75–95) the 'taxonomic' phonologists' concern with segmentation, contrast, distribution and biuniqueness (chapter 4 above) and puts forward the view that phonological description is not based on 'analytic procedures of segmentation and classification' (p. 95) but is rather a matter of constructing the set of rules that constitute the phonological component of a grammar.

5.2 *The sound pattern of English*

Chomsky and Halle's major contribution to phonology, *The sound pattern of English* (1968), is on the one hand an alternative to 'taxonomic' phonemics, and on the other an ambitious attempt to build a description of English phonology on a transformational–generative theory of language. The book (henceforth, as widely, referred to as SPE) begins with a theoretical foundation, arguing that a grammar is a system of rules that relate sound and meaning (p. 3). There are several components of such a grammar, including a phonological component which relates grammatical structures (i.e. grammatically organized strings of morphemes) to their phonetic representations. The heart of SPE (chapters 3 to 5) deals with how such a component of English grammar can be formally expressed.

Chomsky and Halle call attention to numerous alternations in English – what their predecessors would have called morphophonemic rules (section 4.10 above). They classify as 'tense' the vowels in the final syllables of words such as

insane, prostate, explain
obscene, esthete, convene
divine, parasite, divide
verbose, telescope, compose
profound, pronounce, denounce.

Each of the five tense vowels has a corresponding 'lax' vowel, as in

insanity, prostatic, explanatory
obscenity, esthetic, convention

divinity, parasitic, division
verbosity, telescopic, compositor
profundity, pronunciation, denunciation.

Noting the patterns of such alternations, Chomsky and Halle propose various rules which 'tense' and 'lax' vowels in appropriate environments. This means that a word like *convene* can be assigned an underlying form containing a vowel which is lax or tense according to its environment – lax, for instance, before two consonants (as in *convention*) and tense when no suffix is present (as in *convene*). The rules are intended to encompass all the relevant conditioning environments (before CC, before *-ic*, etc.) and include changes to tense vowels such that they are realized as the appropriate long vowel or diphthong. The tense counterpart of [æ] must surface as the diphthong [eɪ] (as in *sane*); the tense counterpart of [ɛ] as the long vowel [i:] (as in *convene*), and so on. The 43 rules finally presented (summarized in SPE chapter 5) are not only complex but include some formal intricacies to do with the abbreviation and ordering of rules (sections 5.4 to 5.6 below). A separate chapter (SPE chapter 8) summarizes and explains the formal apparatus.

Students of the history of the English language will note that the rules of tensing and laxing correspond fairly closely to changes that have taken place in the pronunciation of English. In the fifteenth century, English vowels were subject to a substantial shift known as the Great Vowel Shift. Before this change, for example, the current diphthong [aɪ] in words such as *time*, *wide* and *dine* was almost certainly a long [i:], while the vowel now pronounced [i:] (as in *green* and *meet*) was a long [e:]. Since short (or lax) vowels were not affected in the same way, alternations of the kind mentioned above are largely a consequence of the Great Vowel Shift. It is therefore no coincidence – given the highly conservative conventions of English orthography – that Chomsky and Halle's pairs of tense and lax vowels appear in English spelling as 'long' and 'short' values of the five vowel letters. (In terms of articulation and perception, they are by no means long and short counterparts; see section 10.7 below for consideration of this point in the context of feature systems.)

While Chomsky and Halle are careful not to base their analysis on historical forms – the phonological rules of today's English cannot be justified by appeal to past sound changes – they do include in SPE a chapter on the historical development of the English vowel system (chapter 6) and they do note that 'underlying lexical forms in English contain vowels in pre-Vowel-Shift representation' (p. 332). Elsewhere in SPE, they argue that conventional English spelling is in fact 'a near optimal system for the lexical representation of English words' (p. 49). Their justification for this view – one which is surprising both to those who espouse a phonemic view of phonology and to those who know the struggle of mastering English spelling – is that 'the fundamental principle of orthography is that phonetic variation is not indicated where it is predictable by general rule' (p. 49). Thus wherever speakers know a rule, say that an underlying tense vowel is laxed before the suffix *-ic*, they ought to prefer a spelling convention that presupposes operation of that rule.

The implication that SPE envisages rules applying to segments such as [i] or [o] is actually misleading. Although Chomsky and Halle, and most generative phonologists following them, frequently quote rules containing segmental symbols, they insist that any such symbols are merely convenient shorthand for arrays of features. Thus the symbol [i] is really shorthand for something like

+ syllabic
– consonantal
+ voiced
+ high
..(etc.)..

where the segment is specified as a set of phonetic feature values. A string of segments in comparable notation is sometimes referred to as a matrix, since each segment can be viewed as a set of values entered against the features. The word *deep* [di:p] might be displayed as

d	i:	p
– syllabic	+ syllabic	– syllabic
+ consonantal	– consonantal	+ consonantal
+ voiced	+ voiced	– voiced
– high	+ high	– high
..(etc)..	..(etc)..	..(etc)..

Chapter 7 of SPE gives details of the features, which Chomsky and Halle consider to be the elements of a 'universal phonetic framework' (chapter 10 below). Rules are in principle expressed in terms of these features (as argued by Halle), so that a rule derives one feature specification from another. According to SPE, features are binary at the underlying level (i.e. they take the value + or –) but may have more than two values at the phonetic (surface output) level. (The final chapter of SPE – chapter 9 – does, however, recast feature specifications in a way that has caused major discussion; see section 5.8 below.)

5.3 Basic rule notation in generative phonology

Typically, a phonological rule states that a certain class of segments undergoes a change in some particular environment. For example, a rule may state that obstruents are voiced following any voiced segment. Using the features of SPE (which are listed in appendix 2.2 and further discussed in chapter 10 below), we can write this rule as

(5.3.1) [– sonorant] → [+ voiced] / [+ voiced] __.

The slash comes before the environment specification and the bar on the line indicates the position of the affected segment. A precise but cumbersome reading of the rule is: 'Any segment which is, among other things, nonsonorant is also voiced when standing after any segment which is, among other things, voiced'.

For comparison, here are two rules which state that obstruents are voiced under slightly different conditions:

(5.3.2) [– sonorant] → [+ voiced] / [+ voiced] __ [+ voiced];

(5.3.3) [– sonorant] → [+ voiced] / __ [+ voiced].

Rules refer to classes of segments. Some classes can obviously be specified by a single feature value, such as

sonorants	[+ sonorant]
laterals	[+ lateral]
voiceless segments	[– voiced].

Other classes may require several feature values (again using Chomsky and Halle's features):

$\begin{bmatrix} + \text{ syllabic} \\ - \text{ consonantal} \end{bmatrix}$ = vowels

$\begin{bmatrix} + \text{ syllabic} \\ - \text{ consonantal} \\ + \text{ high} \end{bmatrix}$ = high vowels

$\begin{bmatrix} + \text{ syllabic} \\ - \text{ consonantal} \\ + \text{ back} \\ + \text{ round} \end{bmatrix}$ = back rounded vowels.

Any feature not mentioned immediately to the right of the arrow is assumed to be left intact. Thus by rule (5.3.1), which voices obstruents, a voiceless bilabial plosive becomes a voiced bilabial plosive, a voiceless velar fricative becomes a voiced velar fricative, and so on. The exception to this principle is that there are certain incompatibilities in the feature system. For instance, it is universally impossible for a vowel to be both [+ high] and [+ low], and a rule which makes a vowel [+ high] ought therefore to make it [– low] at the same time, without any need to state this in the rule itself (see section 5.8 below).

It is a principle of generative phonology that phonological rules may refer to grammatical information, specifically that a rule may apply in a particular grammatical domain. The notation includes symbols indicating boundaries, commonly # for the lowest level boundary, ## for the one ranking above it, and so on. By this convention, English morphemes might be separated by #, words by # # and phrases by # # #, e.g.

# #dog# #	dog
# #laugh#ing# #	laughing
# # #the# #laugh#ing# #dog# # #	the laughing dog

Rule (5.3.4) states that consonants are voiceless at the end of a morpheme, (5.3.5) that vowels are high at the end of a word:

(5.3.4) [+ consonantal] → [– voiced] / __ #

(5.3.5) $\begin{bmatrix} + \text{ syllabic} \\ - \text{ consonantal} \end{bmatrix}$ → [+ high] / __ ##.

This notation has the virtue of making it clear that some boundaries are implied by others: a rule that applies in the context –# will also apply in the context –# # or –# # #. An alternative convention uses + for a morpheme boundary, in which case # indicates a word boundary.

The environment of a rule may include several segments, including boundary symbols, e.g.

(5.3.6) $\begin{bmatrix} + \text{ syllabic} \\ - \text{ consonantal} \end{bmatrix}$ → [+nasal]/__ [+nasal]##
(a vowel is nasalized before a word-final nasal segment);

(5.3.7) [– sonorant] → [+ voiced] / ## $\begin{bmatrix} + \text{ cons} \\ + \text{ nasal} \end{bmatrix}$ — $\begin{bmatrix} + \text{ syll} \\ - \text{ cons} \end{bmatrix}$
(an obstruent is voiced if between a word-initial nasal consonant and a vowel).

Other lexical and syntactic information can also be included in the environment, e.g.

(5.3.8) [– sonorant] → [+ voiced] / __ # #]$^{\text{verb}}$

Rule (5.3.8) states that an obstruent is voiced when word-final in a verb; in case this seems improbable, note that some English verbs differ from a cognate noun or adjective in just this way, e.g. *wreath*, *wreathe*, *safe*, *save*.

Classical generative phonology has no symbol for a syllable boundary, and relevant contexts must be specified in other terms – for example, 'in an open syllable' may be equivalent to 'before a single consonant followed by a vowel'. More recently, the need to indicate syllable boundaries has been recognized, and the symbol $ is often used.

Many generative descriptions contain rules which are not worked out in detail. Segmental symbols, including C and V for any consonant or vowel, are often written into rules, e.g.

(5.3.9) C → [+ voiced] / V __ V
(a consonant is voiced between two vowels);

(5.3.10) i → e / __ r C
(the vowel [i] is lowered to [e] before a sequence of [r] plus consonant).

Such rules are informal and it must be assumed that symbols such as C, V, r and so on would be fully worked out in feature notation in a formal description.

The symbol Ø has a semi-formal status as the representation of zero. It appears frequently in the literature but can be regarded as an abbreviation for a feature specification containing [– segment]. The zero symbol appears in rules of deletion and epenthesis or insertion, e.g.

(5.3.11) V → Ø/ / V __ # #
(a vowel is deleted if word-final after a vowel);

(5.3.12) Ø → t / n __ s
(the consonant [t] is inserted between [n] and [s]).

The zero symbol never appears in the description of the environment. Irrelevant components of the environment are simply omitted, so that C__ means 'after a consonant and before anything whatsoever'. But it is sometimes necessary to indicate that something is present, even though its composition is irrelevant. For this purpose dots may be used, or more commonly capital X, Y, Z, W, etc.

(5.3.13a) V → Ø / __ C]^{root} . . .]^{verb};

(5.3.13b) V → Ø / __ C^{root}X]^{verb}.

(5.3.13) gives two versions of a rule stating that a vowel is deleted if it precedes the root-final consonant of a verb. The dots or X specify that the root will be followed by something, perhaps a suffix or an auxiliary element, which falls within the verb but whose composition is of no relevance to the operation of the rule. Actually, notational practice varies: some writers will include boundary symbols whenever they refer to categories such as verb or root, and some seem to prefer to include both opening and closing brackets. The following rules are taken, with some simplifications, from different sources to illustrate notational variety:

(5.3.14) V → Ø/ / V + C __ # #]^{verb}
(within a verb, a suffix of the shape CV loses its vowel if it follows a vowel and stands word-final)

(5.3.15) Ø → ə / $\begin{bmatrix} \# \text{ X C } \_\_ \begin{bmatrix} + \text{ consonantal} \\ + \text{ sonorant} \end{bmatrix} \# \end{bmatrix}$
(a schwa vowel is inserted between two consonants at the end of a word, where the second consonant is a sonorant, e.g. [lm] becomes [ləm], [gl] becomes [gəl]; the rule is formulated so as not to apply across a word boundary, i.e. the two consonants must be within the same word);

(5.3.16) V → Ø / $\begin{bmatrix} [\text{X } \_\_]^{\text{stem}} \text{ V Y} \end{bmatrix}^{\text{verb}}$
(within a verb, a stem-final vowel is elided if before another vowel);

(5.3.17) $\begin{bmatrix} \text{V} \\ -\text{high} \end{bmatrix}$ → [+low]/ __ $\begin{bmatrix} \begin{bmatrix} \text{V} \\ +\text{low} \end{bmatrix}^{\text{stem}} \text{V} \ldots \end{bmatrix}^{\text{verb}}$
(within a verb, a nonhigh vowel is low if it precedes a low vowel which is both stem-final and before another vowel);

(5.3.18) $\begin{bmatrix} + \text{ consonantal} \\ - \text{ coronal} \\ + \text{ high} \end{bmatrix}$ → Ø / __ + [PLURAL]
(a velar consonant is elided before the plural suffix)

All rules dealt with so far are of the format A → B / C __ D, but rules of coalescence and metathesis require special comment. Consider processes such as the coalescence of a vowel and nasal consonant into a nasalized vowel (e.g. [an] → [ã]) or the metathesis of a fricative and plosive (e.g. [sp] → [ps]). Rules expressing such processes apparently do not fit the format. But A → B / C __ D is actually another way of writing CAD → CBD, and this second format is in fact the more general one, allowing us to include more possibilities. In other words, the basic format of a generative rule is one which rewrites one string of symbols as another. For rules of coalescence and metathesis we can retain this format (e.g. ABCD → ACBD); but other rules can be abbreviated into the format we have been using so far, on the understanding that this is a special case of rewriting. Thus a rule of vowel nasalization, with loss of the following nasal consonant, can be written as

(5.3.19) $\underset{1}{\begin{bmatrix} + \text{ syllabic} \\ - \text{ consonantal} \end{bmatrix}}$ $\underset{2}{\begin{bmatrix} + \text{ consonantal} \\ + \text{ nasal} \end{bmatrix}}$ → [+ nasal] $\underset{1}{\text{Ø}}$ $\underset{2}{\text{.}}$

Metathesis of a fricative and plosive can be written as (5.3.20a) or more concisely as (5.3.20b):

(5.3.20a) $\underset{1}{\begin{bmatrix} - \text{ sonorant} \\ + \text{ cont} \end{bmatrix}}$ $\underset{2}{\begin{bmatrix} - \text{ sonorant} \\ - \text{ cont} \end{bmatrix}}$ → $\underset{2}{\begin{bmatrix} - \text{ sonorant} \\ - \text{ cont} \end{bmatrix}}$ $\underset{1}{\begin{bmatrix} - \text{ sonorant} \\ + \text{ cont} \end{bmatrix}}$

(5.3.20b) $\underset{1}{\begin{bmatrix} - \text{ sonorant} \\ + \text{ cont} \end{bmatrix}}$ $\underset{2}{\begin{bmatrix} - \text{ sonorant} \\ - \text{ cont} \end{bmatrix}}$ → $\underset{2}{\text{ }}$ $\underset{1}{\text{ }}$

The following rule of metathesis reverses the order of a glottal stop and consonant when between vowels (e.g. [aʔna] → [anʔa]):

(5.3.21) $\begin{bmatrix}+\text{ syll}\\-\text{ cons}\end{bmatrix}_1 \begin{bmatrix}-\text{ cons}\\-\text{ cont}\\-\text{ distrib}\end{bmatrix}_2 \begin{bmatrix}-\text{ syll}\\+\text{ cons}\end{bmatrix}_3 \begin{bmatrix}+\text{ syll}\\-\text{ cons}\end{bmatrix}_4 \rightarrow 1\ 3\ 2\ 4.$

The following rule says that if a sequence of nasal consonant and plosive occurs between two vowels, then the first vowel is nasalized, the nasal consonant elided and the plosive voiced, e.g. [ampa] → [ãba]:

(5.3.22a)
$$\begin{bmatrix}+\text{ syll}\\-\text{ cons}\end{bmatrix} \begin{bmatrix}+\text{ cons}\\+\text{ nas}\end{bmatrix} \begin{bmatrix}-\text{ son}\\-\text{ cont}\end{bmatrix} \begin{bmatrix}+\text{ syll}\\-\text{ cons}\end{bmatrix} \rightarrow \begin{bmatrix}+\text{ syll}\\-\text{ cons}\\+\text{ nas}\end{bmatrix} \text{Ø} \begin{bmatrix}-\text{ son}\\-\text{ cont}\\+\text{ voic}\end{bmatrix} \begin{bmatrix}+\text{ syll}\\-\text{ cons}\end{bmatrix}$$

or

(5.3.22b) $\begin{bmatrix}+\text{ syll}\\-\text{ cons}\end{bmatrix}_1 \begin{bmatrix}+\text{ cons}\\+\text{ nas}\end{bmatrix}_2 \begin{bmatrix}-\text{ son}\\-\text{ cont}\end{bmatrix}_3 \begin{bmatrix}+\text{ syll}\\-\text{ cons}\end{bmatrix}_4 \rightarrow [+\text{ nas}]_{1\ 2}\ \text{Ø}\ [+\text{ voic}]_3\ 4.$

Of particular interest within generative phonology is the interplay of the notational apparatus and the system of rules taken as an integrated whole. For the sake of simple illustration, the examples given above have been taken in isolation, but in fact any rule will have to be formulated appropriately for a specific language. In languages in which there are no syllabic consonants, for example, the label [+ syllabic] will be adequate to refer to vowels; in other languages the specification may have to include [– consonantal] as well as [+ syllabic]. Moreover, alternative rules may be possible. For example, a single rule that coalesces vowel plus nasal consonant into a nasalized vowel may be better expressed in two rules: instead of

(5.3.23) V N → Ṽ e.g. [an] → [ã]

we might have

(5.3.24) V → Ṽ / __ N e.g. [an] → [ãn]

(5.3.25) N → Ø / Ṽ __ e.g. [ãn] → [ã]

But in postulating rules (5.3.24) and (5.3.25) we are assuming of course that forms will undergo both rules – that (5.3.24) will 'feed' (5.3.25). This raises the question of how rules may interact with each other and of how we might choose between a series of relatively simple interacting rules and a set of more complex but independent rules. We turn to the formalism and its part in evaluating descriptions before going further into questions of rule interaction and rule order later in the chapter.

5.4 Formalism and evaluation

It is possible to distinguish in a very general way between formal and informal approaches to description and explanation of a variety of phenomena. There is a kind of question that asks for the next number in a series such as

1, 3, 6, 10, 15, 21, 28, . . .
or 2, 5, 11, 23, 47, 95, 191, . . .

Those of us familiar with these questions (whether or not we believe they test anything worthwhile) will look for a pattern or rule so that we can generate the next number. If we cannot state a formal rule (say, k = 2j + 1) or at least produce an answer from a tacit understanding of such a rule, we have failed to explain the series.

On the other hand, if we were asked to identify paintings by famous artists, we would expect to adopt a far less formal approach. We might be able to identify a Rembrandt by general similarity with other Rembrandts which we have seen, and by attention to such characteristics as contrast between light and dark, predominance of certain colours, details of the subject itself, and so on. But we are not likely to think of our criteria as formal rules, let alone express them in formal terms.

It is an intriguing question whether these two kinds of task are as different as they seem. If the brain works always with finite possibilities, then the identification of the authorship of a painting may be just as 'rule-governed' as the identification of the next number in a series: it may only be that the rules or procedures are so much more intricate that we are scarcely able to make them explicit. With a series of numbers we deal with reality in a single dimension, as it were; with paintings we have to consider various scales and values, such as colour, brightness, shape and texture, which are integrated in complex ways as design or imagery or style. Whatever the nature of our mental processes and knowledge, it is customary practice to expect relatively formal description and explanation in some fields (such as mathematics and physics) and to expect it much less in others (such as esthetics or the study of art or literature).

In phonology, generative linguists are firmly on the side of a formal approach. The very term 'generative' draws on a mathematical concept of definition by the application of rules or operations. Thus in generative linguistics, a set of rules may be said to 'define' a language by generating all and only the correct possibilities. A language in which every word consisted of one or more occurrences of [m] followed by one or more occurrences of [a] would be defined by a rule that generated any number of [m]s preceding any number of [a]s. Despite the simplicity and artificiality of a language of this kind, it is worth noting that the number of words is infinite, if there is no upper limit on the number of occurrences of [m] and [a]. The rule is therefore powerful, in the sense that it generates an infinite number of possibilities, but also restrictive, in the sense

that it generates only sequences of the language and not impermissible sequences like [aa], [m] or [aaammm]. Indeed, rules are too powerful if they generate not only what is required but also a lot more besides. Hence the predictive or explanatory value of a model of language cannot be equated with generative power: the model needs to be constrained, not open-ended. And one of the challenges facing generative linguistics has been to restrict or constrain its rule-based model of language in principled ways that are appropriate to explain what we find in natural languages.

A concern with formal and explicit description as such is not unique to generative linguists, and it can be argued that the concern itself was inherited from pregenerative North American linguistics (Anderson 1985, p. 316). In general, language is not only amenable to formal investigation but also demands some degree of descriptive formality to convey its true nature. While there may be some value in attempting global characterizations of the phonology of a language, the risk of vagueness and inaccuracy is high. A claim that Dutch or German is a 'guttural' language, for example, means little unless perhaps refined into a statement about the perceptual quality of velar or uvular fricative articulation; likewise a comment that English consonants 'tend to assimilate to a following consonant' again needs to be made more precise, for example by specifying which consonants assimilate, what features are changed in the process and under what conditions. Without such refinement, the comments are tantamount to explaining a series of numbers by saying that each number is 'a lot higher than' the one before it. And refinement and precision bring with them the need for a formal apparatus with which to specify sounds and features and their patterning. What is characteristic of the generative approach, then, is not so much formality and explicitness in themselves but the way in which these goals have been debated and expressed in a rule-based conception of language.

The fundamental reason for formality is the requirement for precision and accuracy. But from this follow further principles, which have been strongly emphasized within generative phonology. First, if the formalism is relatively strict, it limits what can be said. Since models can be too powerful, formal limits are a descriptive strength: the limits make claims about what is possible and therefore make the formal apparatus an expression of a theory of language. If, for instance, there is no limit to the kinds of sounds or rules that can appear in a phonological description, then there may indeed be no reason to constrain the formal apparatus. But if it is true, as most of us believe, that there are limits, then these limits can be expressed or implied by specifying an inventory of features or a set of parameters or a format for rules. Our formal apparatus may then prove to be wrong – if, for example, it turns out to be inadequate for some of the world's languages – but this is precisely what we want, namely that the apparatus makes a claim about the nature of language which can be disproved. If disproved, the claim can be revised and the formal apparatus amended accordingly. In this way, formalism functions as part of the model-building and hypothesis-testing which are characteristic of modern science.

Secondly, a strict formalism of the type favoured by generative phonology provides its own inbuilt measure of what is the simplest and best description.

Halle's point about the use of features in rules (section 5.1 above) is central here. If phonological rules are expressed in ordinary English, with few if any constraints on the wording, it is hard to judge what counts as a simple or plausible rule. But if rules follow a certain format, using features and a limited number of notational devices, then we can measure the complexity of a rule by the complexity of its expression. Here, according to Chomsky (1964), other models of phonology are weaker than the generative model: other phonologies may offer 'descriptive adequacy' but they fail to achieve the 'explanatory adequacy' of a model in which evaluation of the description is inherent in the description itself.

The analogy with explaining a series of numbers may again be helpful. The requirement that such explanation be formulated as a rule implies a framework that both limits the possible answers and provides a measure of simplicity. A rule $k = 2j + 1$ conforms to the format, has explanatory power which can be checked against the series, and can be easily evaluated against an alternative formulation such as $k = 4(j/2 + 1/4)$.

5.5 Abbreviatory devices in rule notation

In previous sections we have touched on two assumptions: that rules may apply in a certain order (section 5.3) and that a rule can be evaluated by counting the number of features in it (sections 5.1 and 5.4). While the arguments for these assumptions are clear enough, the implications are not straightforward. In particular, the notation of orthodox generative phonology includes a number of so-called abbreviatory devices, which have the effect of (partially) amalgamating some rules that come next to each other in the sequence of application. The amalgamated rule then counts more cheaply, by virtue of having fewer features.

Consider, for example, the deletion of /r/ in many varieties of English, where /r/ is not pronounced before a consonant (as in *ear-lobe* or *ear-muff*) or at the end of a word when nothing follows (*ear*) and is retained only before a vowel (*ear-ache*, *my ear is . . .*). The deletion applies in two environments, suggesting two rules:

(5.5.1) $r \rightarrow \emptyset \; / \; \_\_ \; C;$

(5.5.2) $r \rightarrow \emptyset \; / \; \_\_ \; \#\; \#.$

Assuming that these two rules are ordered next to each other, are they really distinct or can we take them as variants of a single r-deletion rule? If the two can be collapsed as

(5.5.3) $r \rightarrow \emptyset \; / \; \_\_ \; \left\{ \begin{matrix} C \\ \# \; \# \end{matrix} \right\}$

then the number of features is clearly reduced, by mentioning /r/ and Ø only once. This abbreviation is legitimate in orthodox notation. It is signalled by the use of BRACES (curly brackets) and applies only to adjacent rules and only where environments can be (partially) combined. (Where it might seem possible to use braces on the left-hand side of a rule, the expectation is that one could achieve the necessary generalization by choice of features; thus instead of bracketing, say, [l] and [r], one should be able to specify non-nasal sonorants.) A condition attached to the use of braces is that the abbreviated rules are taken to be CONJUNCTIVELY ordered. That is, if two rules, collapsed by use of braces, can both apply to a particular string, then both of them *must* apply, one after the other.

Adjacent rules may be similar in a different way if the environment of one is equivalent to part of the environment of the other. Suppose that vowels undergo a certain process both before a single consonant (followed by a vowel) and before certain sequences of consonant, say nasal plus other consonant. The two environments __ C V and __ N C V can be combined as __ (N) C V. (Processes conditioned by environments of this kind are quite common and in most cases are best explained as applying in open syllables, where the nasal consonant does not close the preceding syllable but begins the following syllable; since orthodox generative phonology does not recognize syllable boundaries, it has to formulate the environment in terms of sequences of consonants and vowels.) In Javanese, for instance, /a/ is rounded to a low back rounded vowel (sometimes written /â/) in certain open syllables: the rounding applies in the last two syllables of words such as *râjâ* ('king') and *negârâ* ('country'), and also in *kândâ* ('tell') and *tâmpâ* ('receive'), where the *n* does not close the first syllable but counts as part of the second syllable; on the other hand, rounding does not apply to the first (closed) syllable of words such as *warnâ* ('colour') or *jalmâ* ('human being'). The Javanese rule can be written as

(5.5.4) a → [+ round] / __ ([+ nasal]) C V.

As with the previous abbreviatory device, the assumption is that a rule of this kind is actually two or more rules collapsed into one. The conditions attached to the convention, marked by PARENTHESES or round brackets, are first that the longer rule (including the elements in parentheses) is presumed to precede the shorter, and secondly that the component rules are ordered DISJUNCTIVELY, meaning that once one has applied, any subsequent rules are skipped, whether applicable or not.

In rule (5.5.4) the disjunctivity of the two abbreviated rules is irrelevant, but consider the following rules of stress assignment. Suppose for simplicity's sake that we are dealing with a language in which every syllable is of CV shape and that [+ stress] can be regarded as a value assigned to any stressed vowel. The stress in this language is antepenultimate, i.e.

monosyllables are stressed: 'CV;
two-syllable words have stress on the first syllable: 'CVCV;
words of three or more syllables have stress on the third syllable from the end: 'CVCVCV, CV'CVCVCV, etc.

As a first approximation we might have three rules:

(5.5.5) V → [+ stress] / __ # # (monosyllables)

(5.5.6) V → [+ stress] / __ CV# # (two-syllable words)

(5.5.7) V → [+ stress] / __ CVCV# # (longer words).

If we amalgamate these into one rule, using parentheses, we have

(5.5.8) V → [+ stress] / —— ((CV) CV) # #.

By convention, the expansions of (5.5.8) apply in descending order of size, i.e. in the order (5.5.7), (5.5.6), (5.5.5), and once one of these applies, no other may apply. This is precisely what is necessary to obtain the correct results in this instance. In the case of a two-syllable word, rule (5.5.7) will not apply, (5.5.6) will, assigning stress to the first syllable, and (5.5.5) could apply but should not, as it would assign an additional stress to the final vowel.

A deceptively simple notation in which, for instance, C_1^3 is used to mean 'at least one and not more than three consonants' is actually equivalent to the use of parentheses. Given the formality of generative notation, the conditions that apply to the use of parentheses must be understood to apply to the use of subscript and superscript numbers. Thus C_1^3 is shorthand for (((C)C)C), which will expand into CCC, CC and C, applied disjunctively in that sequence. Further examples of the notation are

C_0^2 two consonants, one consonant or none
V_1^2 two vowels or one
C_1 at least one consonant.

Examples such as the last imply an infinite series of expansions without any principled limit on the maximum number of segments. The notation avoids the problem of having to specify the longest expansion (which should of course be first in the sequence of expansions). This is probably more relevant to syllables than to segments, since the number of syllables per word is likely to be less constrained than the number of consonants or vowels in a cluster or sequence. Hence abbreviations such as $(CV)_1$ or $(CVC)_0$ may be useful in rules that need to skip over an indefinite number of syllables. Anderson (1974, p. 101, appealing to data from Tryon 1970) suggests that Tahitian has just such a rule of stress assignment, in which it would be arbitrary to fix an upper limit on the number of syllables that a word can contain.

A further extension of the parentheses notation is the use of ANGLED BRACKETS to enclose two optional elements that are either both present or both absent. Thus the environments C __ C and VC __ CV could if necessary be combined as < V > C __ C < V >. As with parentheses, the longer expansion applies first and ordering is disjunctive. A more realistic example of angled brackets is a rule such as

(5.5.9) $\begin{bmatrix}\text{+syllabic} \\ \text{< +high >}\end{bmatrix} \rightarrow$ [+stress]/ __ < CV > # #.

The rule states that a high vowel receives stress before CV# # or, if this condition is not met, any vowel is stressed before # #. Disjunctive ordering ensures that final vowels will not be stressed in words that have already received stress on a penultimate high vowel.

It is possible to combine different brackets where appropriate. In some varieties of Indonesian, a vowel is 'tense' if it precedes a consonant plus vowel, or if it precedes a nasal consonant plus consonant plus vowel, or if it is word-final:

(5.5.10) $V \rightarrow$ [+tense] / __ $\begin{Bmatrix}(([+\text{nasal}])\ CV \\ \#\ \#\end{Bmatrix}$

Here the ordering conventions happen to have no relevance, but it is important to realize that they are implied by the notation. Generative phonology hypothesizes that rules that show relevant formal resemblances must be amalgamated and applied in accordance with the conventions. The hypotheses include the claim, for instance, that two rules are disjunctively ordered if and only if they can be combined using braces.

A further notational device is suggested by the existence of complementary rules. A common kind of assimilation simply adjusts a feature to the same value as that of the following segment. In Dutch, for instance, fricatives are as a rule voiceless before voiceless consonants and voiced before voiced consonants: in the plural noun *hoofden* ('heads'), the fricative is voiced to [v] before voiced [d], but in the singular *hoofd* (where the word-final plosive is devoiced), the fricative is voiceless in agreement with the following voiceless plosive. In cases such as these, we may appear to have two rules, one of voicing and one of devoicing:

(5.5.11) $\begin{bmatrix}\text{– sonorant} \\ \text{+ continuant}\end{bmatrix} \rightarrow$ [– voiced]/ __ [– voiced];

(5.5.12) $\begin{bmatrix}\text{– sonorant} \\ \text{+ continuant}\end{bmatrix} \rightarrow$ [+ voiced]/ __ [+ voiced]

But the two rules are actually opposite sides of the same coin and may be combined into a single rule. As with other abbreviatory devices, amalgamation into a single rule amounts to a hypothesis that two related rules count more cheaply in the evaluation system. In this case, the notation allows (5.5.11) and (5.5.12) to be combined as

(5.5.13) $\begin{bmatrix}\text{– sonorant} \\ \text{+ continuant}\end{bmatrix} \rightarrow [\alpha\text{voiced}]/$ __ $[\alpha\text{voiced}]$

The alpha symbol is sometimes referred to as a FEATURE COEFFICIENT and is, technically, a variable ranging over the values + and – (and any other values that may be assigned to a feature, if such there be). The variable must occur at least twice in a rule, and any rule which contains alphas has only two expansions, one in which every occurrence of the alpha is plus, the other with alpha as minus throughout.

The alpha variable has an obvious use in assimilation rules, but the features marked as agreeing in value need not be one and the same feature. Rule (5.5.14) says that obstruents are voiced before sonorants but voiceless before obstruents:

(5.5.14) [– sonorant] $\rightarrow [\alpha\text{voiced}]$ / __ $[\alpha\text{sonorant}]$.

As a further example, rule (5.5.15) states that back vowels are rounded and other vowels are unrounded when before a consonant:

(5.5.15) $\begin{bmatrix}\text{+ syllabic} \\ \text{– consonantal} \\ \alpha\text{back}\end{bmatrix} \rightarrow [\alpha\text{round}]$ / __ [+ consonantal].

The use of a minus sign in front of one alpha allows reference to features which are opposite in value. Thus (5.5.16) and (5.5.17), expressing a dissimilatory process whereby [l] becomes [r] before [l] and [r] becomes [l] before [r], can be abbreviated as (5.5.18):

(5.5.16) $\begin{bmatrix}\text{+ sonorant} \\ \text{– nasal}\end{bmatrix} \rightarrow$ [– lateral] / __ [+ lateral];

(5.5.17) $\begin{bmatrix}\text{+ sonorant} \\ \text{– nasal}\end{bmatrix} \rightarrow$ [+ lateral] / __ [– lateral];

(5.5.18) $\begin{bmatrix}\text{+ sonorant} \\ \text{– nasal}\end{bmatrix} \rightarrow [\alpha\text{lateral}]$ / __ $[-\alpha\text{lateral}]$.

Or consider rule (5.5.19), which says that a word-final [n] is syllabic if it follows a nonsyllabic segment (such as a plosive) but is otherwise nonsyllabic:

(5.5.19) $\begin{bmatrix}\text{+ consonantal} \\ \text{+ nasal}\end{bmatrix} \rightarrow [\alpha\text{syllabic}]$ / $[-\alpha\text{syllabic}]$ __# #.

Where more than one variable is needed, successive letters of the Greek alphabet are used. Assimilation rules often require that segments agree in a number of feature values. In Indonesian, the final nasal consonant of certain prefixes agrees in point of articulation with the following plosive; this is evident in, for example, the agent noun prefix, which is *pem-* before a bilabial, *pen-* before an alveolar, and so on:

bantu (help)	*pembantu* (helper)
duduk (sit)	*penduduk* (inhabitant)
jahit (sew)	*penjahit* (tailor)
guna (use)	*pengguna* (user).

Indonesian *j* is described sometimes as an affricate, sometimes as a palatal plosive, but we assume in any case that the *n* preceding *j* is palatal by assimilation and equivalent to the consonant otherwise written as *ny* in Indonesian. There are thus four points of articulation, which, in the SPE system, can be captured by the features [anterior], [coronal], [high] and [back] (described in appendix 2.2). Hence the feature specifications are:

bilabials (m,b) $\begin{bmatrix} +\text{ anterior} \\ -\text{ coronal} \\ -\text{ high} \\ -\text{ back} \end{bmatrix}$

alveolars (n,d) $\begin{bmatrix} +\text{ anterior} \\ +\text{ coronal} \\ -\text{ high} \\ -\text{ back} \end{bmatrix}$

palatals (ny,j) $\begin{bmatrix} -\text{ anterior} \\ -\text{ coronal} \\ +\text{ high} \\ -\text{ back} \end{bmatrix}$

velars (ng,g) $\begin{bmatrix} -\text{ anterior} \\ -\text{ coronal} \\ +\text{ high} \\ +\text{ back} \end{bmatrix}$

The rule of assimilation must therefore specify that the nasal consonant agrees in each of these four features, as shown in rule (5.5.20):

(5.5.20) $\begin{bmatrix} +\text{ consonantal} \\ +\text{ nasal} \end{bmatrix} \rightarrow \begin{bmatrix} \alpha\text{anterior} \\ \beta\text{coronal} \\ \gamma\text{high} \\ \delta\text{back} \end{bmatrix} / \begin{matrix} \# \\ \_\_ \end{matrix} \begin{bmatrix} \alpha\text{anterior} \\ \beta\text{coronal} \\ \gamma\text{high} \\ \delta\text{back} \end{bmatrix}$

Note that each Greek letter variable is independent of the others: the two alphas must have the same value as each other (+ or −) but need not agree with the other variables, and so on.

In general the question of how rules like these are to be expanded and applied is of no importance, for only one subpart of the rule can apply in any relevant environment. There are, however, some rules which seem to be candidates for the Greek letter notation but which do raise a problem of ordering, namely EXCHANGE RULES. Exchange rules yield an interchange of values (e.g. i → e and e → i) and, to say the least, they are extremely rare. Anderson (1974, pp. 92–7) and Zonneveld (1976) mention examples. One of those quoted by Anderson concerns the formation of plurals in Dinka, a language of the Sudan

in which plurality is indicated by a reversal of the vowel length of the singular form: thus the plural of [pal] 'knife' is [pa:l] 'knives', while the plural of [tʃi:n] 'hand' is [tʃin] 'hands'. The rule must be something like this:

(5.5.21) $\begin{bmatrix} \text{V} \\ \alpha\text{long} \end{bmatrix} \rightarrow [-\alpha\text{long}] / \_\_ \text{ X}]^{\text{noun plural}}$

Such a rule cannot be taken to be an abbreviation of two conjunctively ordered rules, for the second would simply undo the effects of the first. On the other hand, imposition of disjunctive ordering is arbitrary, since it would make no difference which of the two subrules came first, as long as the other subrule was blocked from applying after it. In fact Anderson (1974, p. 94), appealing to the spirit of Chomsky (1967), proposes that rules abbreviated by the use of feature coefficients are a special exception to the principle of rule ordering. The rules apply not sequentially but simultaneously.

5.6 Rule order

In the earliest orthodoxy of generative phonology, rules applied in a fixed order, one after the other. There were exceptions to this principle, namely the special cases signalled in the notation by parentheses and Greek letter variables, which indicated disjunctive or simultaneous application (section 5.5 above); but apart from these well-defined exceptions, rule order was LINEAR, TRANSITIVE and CONJUNCTIVE. The rules of a language could be listed in a numbered sequence; each rule would appear only once in the list; and the output of each rule was the input to the next applicable rule in the numbered order. This early orthodoxy is discussed by Chomsky (1967).

Implicit in these principles of rule order is the assumption that the order must be determined empirically, that the rules of a language take whatever order yields the correct outputs in the most economical way. In other words, order is EXTRINSIC, imposed by the description and not derived from general principles or from the nature of the rules themselves. Indeed, examples have been quoted to show that two dialects might share certain rules but differ in the ordering of them. Sommerstein (1977, pp. 159–61, based on Newton 1972) illustrates the point from Modern Greek. Some dialects share rules which, among other things

1 turn mid vowels into high when next to a low vowel;
2 turn high vowels into semivowels when next to a vowel;
3 turn semivowels into voiced fricatives under certain conditions (e.g. [w] → [v]);
4 delete voiced fricatives between vowels.

The order of these rules does appear to differ among the relevant dialects. In most of the dialects a form such as /aloɣas/ 'horsedealer' is pronounced [aloas],

which suggests that the rules apply in the order given above: the voiced fricative is deleted by (4), but, at this point in the sequence, rule (1) has already been skipped and cannot apply to the mid vowel standing next to a low. In one dialect spoken on Rhodes, however, 'horsedealer' is [alvas]: (4) has applied and the output has then undergone (1), (2) and (3), i.e. [aloas] – [aluas] – [alwas] – [alvas]. Other facts argue that all four rules are present in all the dialects. Their order is therefore crucial.

It has always been apparent, however, that there are difficulties with the orthodox view of rule order. An early attempt to allow some rules to be repeated (ostensibly violating linear transitive order) was the postulation of CYCLICAL RULES. Certain rules were assumed to form a block which could be repeated in a series of CYCLES. In keeping with the generative penchant for constraining the model, only some rules qualified for cyclical application, namely those which were both deep (i.e. 'early' in the total set of ordered rules) and sensitive to syntactic information. Hence successive cycles are not arbitrary but correspond to increasingly larger syntactic domains. A set of cyclical rules might apply first of all within morphemes; on the second cycle the same rules would apply again within words; on the third within phrases; and so on. Thus the cycle was not a means of repeating any rule at random, and linear conjunctive order was still the norm. In SPE it is only the stress rules of English which are cyclical, and other rules are postcyclical (Chomsky and Halle 1968, chs 2, 3, esp. pp. 15–24; some details can be found in the treatment of English prosody in chapter 9 below). It has been suggested that stress is also assigned cyclically (either entirely or partly) in other languages, including Russian (Halle 1973), Japanese (McCawley 1968) and Spanish and Arabic (Brame 1974). Brame goes so far as to hypothesize that stress rules are cyclical in all natural languages.

An example of cyclical rule application other than stress assignment is given by Harms (1968, pp. 99–100). In Komi, a Finno-Ugric language spoken in northern Russia, the vowel [ɨ] is inserted between consonants to avoid clusters of three consonants. But in a word such as *pukśini* the vowel is inserted between *ś* and *n*, whereas in *vundiśni* the vowel is inserted between the *d* and *ś*. The correct form can be predicted, according to Harms, if the insertion rule is applied cyclically. The structure of the two words can be represented as

puk + ś + nɨ i.e. [[[puk][ś]] nɨ]
vund + ś + nɨ i.e. [[[vund][ś]] nɨ]

and the insertion rule can be written as

(5.6.1) Ø → ɨ / [X C C __ C Y].

Now the rule will 'search' for a string that meets its requirements, working upwards from the smallest constituents. On the first cycle, searching within the innermost brackets, the rule will fail to apply. On the next cycle, the innermost brackets are now ignored, and insertion applies to the three consonants

within the string [vund ś]; but it is not applicable to [puk ś] since the CC __ C environment is still not to be found. On the next cycle [vund ɨ ś nɨ], having undergone the insertion rule, no longer has a CCC sequence; but [puk ś nɨ] does now trigger insertion, at the appropriate point in the string. Cyclical treatment of other segmental phenomena in North American Indian languages was proposed by Kisseberth (1972, for the language Klamath) and Kaye and Piggott (1973, to account for palatalization in Ojibwa).

A very simple summary account of conjunctive and disjunctive order as handled in the early days of generative phonology can be found in Schane (1973, pp. 89ff.). Chomsky and Halle's own treatment of English stress rules, accompanied by some discussion of the ordering conventions, is in chapter 2 of SPE. For more general evaluation of the hypotheses themselves and their validity, see Anderson (1974, chs 6 and 7) and Sommerstein (1977, ch. 7).

Even with these various exceptions or qualifications, the principle of linear transitive order has faltered in the face of various examples of ORDERING PARADOXES (Anderson 1974, pp. 141ff.; Sommerstein 1977, pp. 174–6). In Icelandic, for instance, there are two rules, one of which is an umlaut rule converting /a/ to a front rounded /ö/ before an /u/ in the following syllable, the other an elision rule deleting unstressed vowels in certain environments. Slightly simplified, the two rules are

(5.6.2) a → ö / __ C_o u;

(5.6.3) $\begin{bmatrix} V \\ -\text{stress} \end{bmatrix}$ → Ø /C __ C # V

Thus in the nouns *jökull* ('glacier') and *jötunn* ('giant') the first vowel is an underlying /a/ which has become /ö/ because of the /u/ in the following syllable. Now the dative form of 'glacier' is *jökli*, from underlying /jakuli/: the /u/ triggers assimilation of the /a/ in the first syllable but is then deleted by the elision rule. Thus the two rules seem to apply in the order given above. But the dative plural of 'gods' is *rögnum*, from underlying /raginum/. Here – and in comparable forms such as *kötlum* ('kettles'), from underlying /katilum/ – the rules must apply in the reverse order: the unstressed /i/ is elided, which then allows the /u/ of the last syllable to trigger rounding of the preceding /a/.

Paradoxes such as these prompted a number of suggestions about principles of rule order. One proposal envisaged PARTIAL ORDER: rules would be unordered and could apply whenever and wherever their conditions were met, but some of the rules might be specified as preceding certain others, or as blocking the subsequent application of certain others. Or most rules might fall into an ordered set, but some, termed PERSISTENT RULES or ANYWHERE RULES (Chafe 1968; Anderson 1974, p. 191), would be capable of applying as often as they could. Or, under a principle known as LOCAL ORDER, the order of precedence might be specified only for pairs of rules at a time. (See Sommerstein 1977, pp. 176–88, for an overview.)

5.7 Functional considerations

Debate about rule order led to reconsideration of functionality in language. The question arose whether rule order might not in fact be determined by functional or natural principles – whether rule order might be INTRINSIC, i.e. determined by the nature and function of the rules themselves.

In 1968 Kiparsky had already drawn attention to the effects of alternative orders and had distinguished between FEEDING and BLEEDING. If two rules (call them A and B) are such that A generates forms which will undergo B, then A feeds B. If the order of these two rules is reversed (nonfeeding or counterfeeding order), there will be apparent exceptions to B, since A now generates forms that escape the effects of B by virtue of the ordering. Assuming that language is characteristically regular and averse to exceptions, feeding order seems more likely or more natural than counterfeeding. To take a simple example, rule (5.7.1) feeds (5.7.2), since it creates additional occurrences of /r/ to undergo (5.7.2):

(5.7.1) $l \rightarrow r \; / \; — \; \# \; \#$

(5.7.2) $r \rightarrow [-\text{ voiced}] \; /— \; \# \; \#$

It seems unlikely that the order of these two rules would be the reverse. Counterfeeding would mean that those occurrences of /r/ which resulted from (5.7.1) – and only those – would remain voiced in word-final position, violating the pattern implied by (5.7.2).

If two rules (C and D) are such that C robs D of some of its inputs, then C bleeds D. If the order of these two rules is reversed (nonbleeding or counterbleeding), then the application of D will be maximized instead of constrained. Here counterbleeding seems the more natural order. For example, rule (5.7.3), which raises /a/ to /e/ before any palatal consonant, bleeds (5.7.4), which nasalizes the low vowel before any nasal.

(5.7.3) $a \rightarrow e \; / \; — \begin{bmatrix} - \text{ anterior} \\ + \text{ coronal} \end{bmatrix}$

(5.7.4) $a \rightarrow [+ \text{ nasal}] \; / \; — \; [+ \text{ nasal}].$

Bleeding order means that an /a/ standing before a palatal nasal is raised to /e/ and then fails to undergo low vowel nasalization. This seems the most plausible state of affairs, given that rule (5.7.4) applies only to /a/ and therefore would not apply to any occurrence of /e/, whether generated by (5.7.3) or not. The reverse (counterbleeding) order would mean that /a/ standing before a palatal nasal would be nasalized and then raised to become /ẽ/, violating the general pattern that vowels other than /a/ are not nasalized before nasal consonants.

Feeding and bleeding are related to the notions of TRANSPARENCY and OPACITY (Kiparsky 1971). A rule is transparent if its effects are obvious from the phonetic forms of a language. Suppose a language has a rule that underlying word-final [o] becomes [u]. If the language has no instance at all of word-final [o], no instance of [u] other than word-finally, and no instances of word-final [u] other than those derived from [o] by this rule, then the rule is as transparent as can be. If on the other hand the language has some instances of word-final [o] that somehow escape the effect of the rule, some instances of word-medial [u] and even instances of final [u] which are not derived from [o], then the rule is highly opaque. Many rules will of course fall between these two extremes. In English, the reduction of unstressed vowels to [ə] is relatively transparent, at least in varieties such as RP: there are few if any occurrences of unreduced vowels in unstressed syllables, and arguably few instances of [ə] other than those derived by the reduction process (although it is a controversial question whether the [a] in the final syllable of words such as *carrot*, *summon* or *opal* is in any sense derived from a full vowel). On the other hand, what Chomsky and Halle (1968) call the laxing of vowels is relatively opaque. The generative treatment of English predicts, for example, that a vowel will be 'lax' before a consonant cluster (section 5.2 above): hence the change of vowel in e.g. *mean*, *meant*, *sleep*, *slept*, *wide*, *width*. But there are certainly instances of 'tense' vowels before clusters (*fiend*, *heaped*, *pint*, *heights*) and some 'lax' vowels before clusters are not derived from 'tense' vowels (*dent*, *adept*, *crypt*, *lint*).

Kiparsky's discussion of feeding and bleeding was actually in a historical context. He observes that over time 'rules tend to shift into the order which allows their fullest utilization in the grammar' (1968, p. 200), and he quotes instances of languages in which rules have evidently been reordered in line with this tendency. In other words, the historical development of languages seemed to favour feeding and eliminate bleeding. Other historical tendencies have been noted: in a study of Spanish, Harris (1973) suggests that rules tend to shift into the order that favours PARADIGMATIC UNIFORMITY, i.e. rules will occur in whatever order reduces irregularity in the morphology of the language. In Spanish, some verb paradigms are not regular: note the alternation of *c* and *g* in

hacer	[aθer]	to do
hago	[aɣo]	I do
hacemos	[aθemos]	we do.

Now, nonuniform paradigms such as these are, as Harris puts it, a 'vanishingly small minority of Spanish verbs', and it seems that many verbs which once had variable stems have been made regular by the reordering of rules. The stem-final consonant of *cocer* ('to cook'), for instance, must once have appeared as an affricate in some forms of the verb and as a velar plosive in others. In modern Spanish, however, the stems end consistently in [θ] (or [s] in much of the Spanish-speaking world) and it is possible to explain this regularization as the result of reversing the order of two particular rules. But Anderson (1974, p. 208) points out that natural principles may conflict with

each other. He points to the SELF-PRESERVATION of rules, noting that counter-bleeding may be natural where bleeding order would mean that the first rule would actually be lost from the language. But these various historical tendencies are no more than that: they do not preclude exceptions and it is clear, for example, that notwithstanding paradigmatic uniformity, languages may tolerate a high degree of morphological irregularity, and that notwithstanding self-preservation, rules do sometimes disappear from languages (Anderson 1974, pp. 209–18; Kisseberth 1973; Thomason 1976).

While some of this discussion in the 1970s seemed to concentrate on formal mechanisms, attention returned from time to time to functional goals or targets. It was noted, for instance, that rules which appear formally unrelated may nevertheless serve a common functional target, such as elimination of consonant clusters, preservation of distinctiveness or maintenance of a generalized stress pattern.

Kisseberth (1970) argued that a number of rules in Yawelmani (a language of California) had the net effect of severely constraining consonant clusters: 'There are a variety of phonological processes which, it may be said, "conspire" to yield phonetic representations which contain no word-final clusters and no triliteral clusters' (1970, p. 293). Studies of RULE CONSPIRACIES, as they came to be called, included one of the Australian language Yidiny, in which stress and vowel length are subject to intriguing constraints (Dixon 1977). Briefly, long vowels cannot occur in adjacent syllables, and in words with an odd number of syllables, at least one even-numbered syllable must contain a long vowel. Stress falls on the first syllable containing a long vowel (or on the first syllable if all the vowels in the word are short); and, counting outwards from this stressed syllable, stress is also assigned to every even-numbered syllable. For example:

yatjí:rringál
wúngapá:tjinyúnta
tjámpulángalnyúnta.

There are various rules, including even some determining the sequence as well as the forms of affixes, which conspire to maintain the phonotactic constraints. Dixon concludes that the details of affixation and vowel length 'must surely indicate that the development of Yidiny morphology has been in part oriented to the language's overriding phonological targets – that every long vowel should occur in a stressed syllable, and that stressed and unstressed syllables should alternate in a phonological word' (1977, pp. 33–4).

In fact functional targets of this kind are not necessarily captured in a subset of the rules. It can be argued (Kiparsky 1972, p. 216) that English tends to avoid repeating /l/ or /r/ within the same word, but that this is revealed in a variety of phenomena, including the general phonological patterning of words as well as morphological alternations that can be expressed in rules. Consider, first, words such as *prattling*, *sprinkling*, *trampling*, *trickling*, *fluttering*, *glimmering*, *glittering*, *spluttering*. These may contain a cluster containing /l/ and a cluster containing /r/, but not two containing /l/ or two containing /r/. There

are few if any words (and certainly none of this semantic type) that have a shape such as *flickling* or *sprittering*. Secondly, an apparently quite different phenomenon in English is that while many adjectives end in *-al* (*educational*, *occasional*, *cultural*, *dental*, *natural*), *-ar* appears where there is an /l/ in the stem (*cellular*, *circular*, *vulgar*, *lunar*, *alveolar*). This constraint – which actually reflects a pattern of Latin – is not absolute in modem English (cf. *laminal*, *laminar*) and in any case loses some of its force in those varieties of English that no longer pronounce final *r*, but nevertheless seems to tend in the same direction as the patterning of words such as *flickering* and *sprinkling*. Thirdly, it may be noted that while *-al* can also mark nouns in English (*betrayal*, *burial*, *dismissal*, *denial*) there are no such nouns with stems containing /l/ (such as *applial*, *dispellal* or *recoilal*). Conspiracies and functional targets are a problem for a model of phonology that relies on formal devices such as bracketing to unite or relate rules. Indeed, Sommerstein takes the Yawelmani example and others like it as evidence for the traditional recognition of phonotactic constraints as a separate component of phonological description (1977, pp. 194–9).

Kisseberth (1973) had also noted that phonological rules often seemed to operate not according to some arbitrarily imposed order but in a way that was sensitive to the consequences of rule interaction. One of his examples is from Dayak (a language spoken on the island of Kalimantan), as reported by Scott (1964). In Dayak vowels are nasalized following a nasal consonant, e.g.

[mãta]	eye
[nãŋãʔ]	straighten
[nãŋgaʔ]	put up a ladder.

Optionally, a voiced plosive following a nasal can be deleted – but this rule does not feed vowel nasalization. Hence 'put up a ladder' may be pronounced [nãŋgaʔ] or [nãŋaʔ], but not [nãŋãʔ]. This is in one sense unnatural, since we would expect feeding order, making vowel nasalization more transparent. But it also makes functional sense, since the lack of nasalization is what keeps 'put up a ladder' distinct from 'straighten'. In other words, vowels are nasalized after a nasal consonant, provided that the nasal is not derived from a cluster of nasal and voiced plosive. Following Kisseberth's formulation, a constraint of this kind is known as a GLOBAL CONSTRAINT or TRANSDERIVATIONAL CONSTRAINT, as it makes reference to derivational history, carrying out, so to speak, a check on the effect of rules.

As Kiparsky puts it (1972, p. 217), phonological rules tend to avoid the universally complex and to maintain what is distinctive in the language. All languages, for instance, put limits on the clustering of consonants. Some, like Japanese and Polynesian languages, allow few or none at all; in Japanese, for example, no consonant cluster of any kind is tolerated word-initially or word-finally, and word-medial clusters are restricted to lengthened plosives and nasals plus plosives. Other languages, like English, are far more tolerant of consonant clusters but are nevertheless prone to processes of simplification: many speakers of English will elide the bracketed consonants in e.g.

Did you sen(d) my letter?
They kep(t) quiet.

But distinctiveness is not ignored in such processes. The lengthening of nasals before voiced plosives in English is such that the [n] in *sen(d)* may still be significantly long, even when the [d] is dropped, and therefore distinct from the shorter [n] that would signal a following [t]. Of course, distinctiveness is not an absolute requirement, and there is ample evidence that distinctions do sometimes disappear from a language – modern English has, for example, lost the distinction between *ail* and *ale* or *hail* and *hale*. But sociolinguistic studies suggest that speakers may be less likely to apply elision where a crucial distinction is lost. Thus in some varieties of English, elision of the final consonant of *fist* is extremely common – and nothing is really lost, for there is no potential confusion with a word such as *fiss*. Elision of the [t] in *kept* is less frequent: here the [t] is a signal of past tense and perhaps therefore more likely to be retained (although the change of vowel from *keep* serves to maintain distinctiveness). But elision is even less frequent in a form such as *passed* (or *past*), where without the [t] the form is phonetically indistinguishable from *pass*. If it is legitimate to speak of a rule of consonant elision applying to these forms, it is a rule constrained not arbitrarily by its priority in a sequence but in its frequency of application and by its effect on communicativeness (Kiparsky 1972, p. 197; Wardhaugh 1986, p. 178–81). A useful review of rule order and feeding and bleeding, with copious examples, can be found in Kenstowicz (1994, pp. 90–100).

5.8 Naturalness and markedness

Chomsky and Halle begin chapter 9 of SPE with an honest if irritating admission that they are dissatisfied with their treatment of features in the book. They point out that the use of features is intended to provide an inbuilt evaluation of naturalness. Generative phonology implies, for instance, that a natural class of sounds will be characterized by relatively few features. Indeed, the fewer the features needed, the more natural the class of sounds: hence obstruent consonants (which can be characterized simply as [–sonorant]) constitute a more natural class than, say, voiced consonants other than laterals (which might require the specification [+voiced, +consonantal, –lateral]). But SPE's approach to such evaluation is, in Chomsky and Halle's own words, 'overly formal' (1968, p. 400). That is, merely to count the number of features overlooks the 'intrinsic content' of the features. Thus the feature specification [–voiced, –sonorant] (= voiceless obstruents) indicates a more common category of description than, say, [–voiced, +nasal] (= voiceless nasals), yet both need only two features. A similar point can be made about rules, undermining Halle's contention about simplicity (section 5.1 above). Even when expressed in features, the simpler rules do not always appear simpler.

For reasons such as these, Chomsky and Halle proposed that feature values be revised to clarify the extent to which certain rules or combinations of features were expected or natural. They appealed to the terms MARKED and UNMARKED, which had been used by some European phonologists (section 11.6 below) to refer to phonemes which showed the presence or absence of a particular feature. In this usage, voiced phonemes might be described as 'marked' by voicing, in opposition to 'unmarked' voiceless phonemes which lacked the feature. But the unmarked member of an opposed pair was often the one to appear in a position of neutralization (section 4.9 above), and the term 'unmarked' sometimes carried the sense of 'neutral' or 'natural' (what the computer-literate might call the 'default value'). This concept has not been confined to phonology and it is sometimes said, for example, that in an adjective pair such as 'long' and 'short', 'long' is the unmarked term because it is the one used in neutral contexts such as questions. (The question 'How long is that string?', without stress on 'long', need not imply that the string is either long or short, whereas the question 'How short is that string?' does imply that the string is short; hence the choice of 'short' in this context is 'marked'.)

Now strictly speaking, the feature system of SPE is incompatible with the notion of markedness: if features are binary (having only the two values, + and –) then there is no room for a third value, 'unmarked'. In fact, phonologists have experimented with such possibilities, abandoning the binary assumption and allowing three values. For instance, if English /m/ is [+labial] and /n/ [–labial] (among other things, of course), we might allow that the nasal consonant of the prefix *in-* is [0labial], meaning that it is unspecified for this feature: here the nasal consonant takes the feature value of the following consonant and will be [+labial] in e.g. *impossible* and *impertinent*, but [–labial] in e.g. *indecent* or *intolerable*. But Chomsky and Halle kept a binary system, while still attempting to exploit a concept of markedness. They proposed that the binary values of underlying features should be 'marked' and 'unmarked' (abbreviated as *m* and *u*) instead of + and –. These new values would reflect expectedness or naturalness, and would be converted into + and – by UNIVERSAL MARKING CONVENTIONS. Thus if it is more usual or natural that sounds are voiced, a universal convention will specify that [*u*voiced] → [+voiced]. In fact the marking conventions are not quite as simple as this, and many of them are sensitive to context. It is assumed, for instance, that [+voiced] is the natural status of vowels, since they are voiced in most languages and vowel qualities are less audible in voiceless vowels; but plosives are more likely to be [–voiced], since it is physiologically easier to switch off voicing during occlusion (see STOPS in section 2.12 above). To allow for this, a marking convention may specify that [*u*voice] is interpreted as [–voiced] in obstruents, but otherwise as [+voiced]. Similarly, since it seems to be the case that consonant followed by vowel is a natural syllabic structure, another of Chomsky and Halle's marking conventions specifies that [*u*vocalic] is [+vocalic] following a consonant. Universal implications are also incorporated into the conventions, including some which simply reflect the incompatibility of certain values, e.g. [+low] → [–high].

Chomsky and Halle add to these conventions the concept of LINKING (1968, pp. 419ff.), which allows the marking conventions to monitor phonological

rules. In effect, marking conventions are not only a set of initial interpretations, applying before phonological rules go to work, but conditions on the output of rules, so that they may also be triggered by appropriate rules. This is a way of simplifying some rules (and hence enhancing their naturalness) by omitting from them details which can be tidied up by the application of relevant marking conventions.

Chomsky and Halle's concept of markedness has been rejected by most of their successors, on the grounds that it still fails to do justice to naturalness. Concern with naturalness has proved a strong motive in recent phonology, so much so that two 'schools' of phonology have enshrined the term in their titles (sections 11.10 and 11.11 below). Classical generative phonology is, however, less famous for its regard for naturalness than for the degree of abstractness which it allows in phonological analysis.

5.9 Abstractness

Orthodox generative phonology is mentalist, in that it implies mental storage of underlying representations which are converted into surface representations by the application of rules. Chomsky and Halle speak of 'mental construction' by speaker and hearer (1968, p. 14). And in connection with access to underlying representation in the process of reading aloud and with the development of such representation in children's acquisition of language, they refer to the 'fundamental importance of the question of psychological reality of linguistic constructs' (1968, pp. 49–50; see also Chomsky 1964, chs 1, 5, and 1968, for Chomsky's views on the relationship between linguistics and psychology). Much of the early argument for generative phonology (in Chomsky 1964, for instance) was devoted to showing that a traditional phonemic transcription was an unjustifiable level of representation, intermediate between underlying and surface representations. The new underlying level (termed 'systematic phonemic') corresponded to the speaker's storage of phonological representations, while the surface level ('systematic phonetic') remained comparable to a traditional phonetic transcription of the speaker's utterances.

Underlying representation was now 'deeper' or 'more abstract' than a conventional phonemic transcription and could be as abstract as the phonological rules would allow. Thus the underlying form of the morpheme common to the English words *telephonist* and *telephonic* should be such that the appropriate surface vowels can be derived by rules. For convenience (and following the example of most generative phonologists) let us represent the underlying form in segments rather than features as *tElEfOn*: this form now depends for its validity upon the fact that English has phonological rules deriving unstressed [ə] from underlying E in appropriate environments (as in the first syllable of *telephonist*), and [ɒ] from underlying O (again in appropriate contexts, such as in the syllable preceding the suffix *-ic*), and so on. The question that then arose was whether there were principled limits on this strategy of description.

Thus a single form may underlie *south*, *south(ern)* and *sou'(west)*, provided that English includes rules to voice a dental fricative in appropriate places and to delete it in others. But these rules might be regarded as *ad hoc*, devised not to reflect general processes but purely to cater for one or two words. (Note that the final dental fricative can be dropped in *north* and *south* but not in other words such as *mouth*, *birth* and *hearth*.) And if this case seems worrying, what of *go* and *went*? If *went* is grammatically the past form of *go*, could it be derived from *go+ed* by application of rules that turn the underlying initial consonant into [g] or [w] according to context, an underlying vowel into [oʊ] or [ɛ], and so on?

In fact generative phonology recognized this problem quite early, and various restrictions on abstract analyses were formulated. Totally abstract segments were ruled out, for example. This meant that both underlying and surface representations were expressed in standard features, and it was not considered legitimate to postulate abstract features or segments that had no genuine phonetic value. Postal (1968) makes this point in the form of the NATURALNESS CONDITION, which states that a (systematic) phonemic representation implies identical phonetic representation unless the phonological rules determine otherwise. In other words, an underlying representation must be such that it would surface as a pronounceable item in the language without the intervention of any rules. The condition forbids any totally abstract segment that is phonetically invalid until altered or fleshed out by the rules of the language.

A further early proposal was to exclude ABSOLUTE NEUTRALIZATION. (Kiparsky's paper on this subject circulated from 1968 but was published only in 1973.) This exclusion meant that if two segments were distinguished at the underlying level, then they must also be distinct in at least some contexts in surface representation. It should not be possible for phonological rules to turn all occurrences of both segments into identical surface segments. Consider, for example, those varieties of English in which the distinction between voiced and voiceless /w/ has disappeared, so that there is no longer any distinction in pronunciation between *which* and *witch* or *whale* and *wail*. Historically, these versions of English have undergone a sound change which has indeed absolutely neutralized the distinction. There would be no justification for postulating two underlying segments, one voiced and one voiceless, for there would then have to be a rule that turned voiceless semivowels into voiced, to ensure the output of forms as pronounced. In short, we need no underlying distinctions that have no phonetic reflex whatsoever. Of course the constraint on absolute neutralization does not exclude the possibility that distinctions are neutralized under some conditions, or that segments are radically altered by rules; it does require that the distinction survive in surface representations in *some* way under at least *some* conditions.

Limitations such as these still left room for a degree of abstractness in underlying forms that many phonologists found alarming. Indeed, the potential to offer abstract analyses was defended as a virtue of generative phonology, and SPE proposed that English had, among its underlying segments, a front rounded vowel /œ/ (which surfaces as the diphthong [ɔɪ] as in *coin*, SPE pp. 191–2) and a velar fricative /x/ (which never appears on the surface but triggers certain

changes in adjacent segments before it is deleted; SPE pp. 233ff.). The reader will notice that neither of these two segments is readily pronounceable by most speakers of English, and the postulation of an English velar fricative became something of a cause célèbre.

Chomsky himself chose his analysis of the word *righteous* to illustrate the possibility that surface structures might be quite surprisingly remote from what underlay them. Pointing to examples such as

expedite	expeditious
ignite	ignition
delight	delicious

Chomsky suggests that we might expect the adjective *righteous* to follow the same pattern, i.e. *ritious*, rhyming with *delicious*. The actual form *righteous* is in fact unexpected in two ways: it shows [tʃ] instead of [ʃ], and [aɪ] instead of [ɪ]. Now there are other forms which show [tʃ] where [ʃ] might be expected, e.g.

suggest	suggestion
Christ	Christian.

These apparent exceptions can be explained by the presence of the fricative [s] before the [t]. The general rule converting [t] to [ʃ] before the relevant suffixes can be modified to ensure that [t] preceded by a fricative becomes [tʃ] and that [t] otherwise becomes [ʃ]. If the underlying form of *righteous* is assumed to contain a fricative preceding the [t], it will undergo this rule.

There are also instances in English where a velar consonant triggers a change in a preceding vowel and is then deleted. Compare

paradigm	paradigmatic
resign	resignation.

In the generative treatment of English, these forms are assumed to contain an underlying [g] which is preserved in the suffixed forms but is lost before the word-final nasal after conditioning a change of the preceding vowel. If the fricative just postulated in *righteous* is now taken to be a velar fricative, then the rules applying to *paradigm* and *resign* can be extended to *right(eous)*. The various rules of English to which we have now referred will derive the initially unexpected pronunciation. By postulating an underlying velar fricative in *righteous*, we make the form accessible to rules which will not only generate the correct diphthong and affricate but also delete the fricative into the bargain. Moreover, the rules applying to this form have not been invented specially or arbitrarily for this purpose but are already required elsewhere in the description of English phonology. In generative terminology, they are 'well-motivated' rules.

Acknowledging that a single example such as this may be less than convincing, Chomsky nevertheless claims that careful investigation of sound structure 'shows that there are a number of examples of this sort, and that, in general, highly abstract underlying structures are related to phonetic representations by a long sequence of rules . . . Assuming the existence of abstract mental representations and interpretive operations of this sort, we can find a surprising degree of organization underlying what appears superficially to be a chaotic arrangement of data' (1968, p. 36). Thus within the early orthodoxy of generative phonology, a high degree of abstractness, within an explicitly mentalist perspective, was regarded as a cornerstone. Sommerstein reviews the early debate about abstractness in some detail and lists major references (1977, pp. 211–25); Lass (1984, ch. 9), Roca and Johnson (1999) and Gussenhoven and Jacobs (2005) give a thorough overview; and Kenstowicz (1994, pp. 103–14) gives a useful account of the debate about alternants and underlying forms. Not surprisingly, the permissible extent of abstractness remained a matter for discussion and became a key feature of modifications to generative phonology, which are reviewed in chapter 11 below (sections 11.10 onwards). The approach to phonology represented by SPE, taken as a formal apparatus of description, scarcely survived the 1970s; but the concept of a generative model of phonology and the assumption that theory must be expressed in explicitly formal terms amount to a still powerful tradition. Many phonologists still proclaim themselves generativists, and, in that sense, the spirit of SPE lives on.

Exercises

1 The suffixes *-ic* and *-ity* are said to trigger vowel laxing in the preceding syllable in English: compare *mania*, *manic*, *phone*, *phonic*, *sane*, *sanity*, and *obscene*, *obscenity*. Find as many English words as you can containing these suffixes, noting any counterexamples.
2 Review the reasons why Chomsky and Halle proposed an underlying velar fricative in the English word *right*. Why did they not also propose one in *night* and *delight*? What, in general, is the justification for postulating abstract underlying forms?
3 What do you understand by a 'plausible phonological rule'? Give examples of plausible and implausible rules.
4 Write formal rules to express the following processes; if you can see more than one way to formalize a process, note the possibilities and consider reasons for and against alternatives.

   a. fricatives are voiced between vowels
   b. sibilant fricatives are voiced before nasal consonants
   c. obstruents are voiceless if word final
   d. obstruents are voiceless before voiceless consonants and voiced before voiced segments
   e. low vowels are nasalized between nasal consonants
   f. vowels are tense at the end of a morpheme
   g. alveolar consonants are palatalized before high front vowels
   h. any vowel is deleted after any other vowel
   i. a sequence of any vowel followed by [r] is metathesized before a plosive
   j. a sequence of [a] followed by [i] becomes [e]

5 Explain the uses of the notion of markedness in linguistic description.
6 What is naturalness in phonological description? Can some languages be said to have more natural phonological systems than others?
7 What are the principal differences between a conventional phonemic description of a language and a generative description?
8 What are the reasons for using formal notation in phonological description?

6 The Anatomy and Physiology of Speech Production

This chapter provides a comprehensive anatomical background to the book's account of speech sounds. The first two sections set the scene for a technical account using the conventions of anatomical description (6.1 and 6.2).

The bulk of the chapter reviews the various organs of speech in a logical order, moving from the broad underlying structures and functions of the nervous system and respiratory system to the details of specific articulators such as tongue and lips. Given the complex functions of the larynx in speech, the section dealing with the larynx is followed by a separate section on how the larynx functions in phonation. The sections are:

- the nervous system (6.3)
- the respiratory system (6.4)
- the larynx (6.5)
- phonation (6.6)
- the pharynx (6.7)
- the velum and the nasal cavity (6.8)
- the oral cavity (6.9)
- the tongue (6.10)
- the lips (6.11)
- the mandible (6.12).

6.1 Introduction

In chapter 2 we outlined the speech production process from a functional perspective with sufficient detail to allow us to describe the speech sounds of language, but deliberately avoiding much discussion of the underlying technical detail. In this chapter we now provide a more technical examination of the anatomical and physiological processes of speech production. Some readers may choose to skip this and the ensuing chapter on speech acoustics, but for others, and especially those whose interests lie in experimental phonology and phonetics, speech and hearing science, communication disorders, cognitive science, artificial intelligence and speech technology, these two chapters provide an essential