5.1 Naturalness

In Chapter 4 it was seen that simplicity, as measured by feature counting, has played a large role in the development of generative phonology. Since the early years of this theory of phonology, there has been a noticeable shift away from simplicity and feature counting. Instead, phonologists have addressed themselves to the naturalness of phonological properties. Thus there is not only concern with what is simple (that is, general, noncomplex), but also concern for what is natural or plausible in a phonetic sense (see 3.4.4). Certain aspects of phonology are not necessarily (or exclusively) simple, but are rather (or in addition) natural. As a result, these aspects are frequently attested in language after language. The new concern is to be sure that what is "natural" is formally revealed as "simpler" than what is "unnatural." Where a natural property of a phonological system is, by the criterion of feature counting discussed in Chapter 4, not revealed to be simple, the evaluation metric is assumed to be wrong and must be revised (see Chomsky and Halle, 1968, Ch. 9).

5.1.1 Natural Classes

In order to demonstrate the potential difference between simplicity (generality) and naturalness, let us return to the notion of natural classes, mentioned briefly in 4.3.1.2. Two or more segments are said to constitute a natural class if fewer features are required to specify the class than to specify any one member of the class. Consider as an example the class of voiceless stops in English. To specify the class /p, t, k/, three features are required: [−voice, −cont, −del rel]. On the other hand, the following feature matrices are required in order to uniquely specify /p/, /t/, and /k/ individually:

/p/ /t/ /k/
[−voice] [−voice] [−voice]
[−cont] [−cont] [−cont]
[−del rel] [−del rel] [−del rel]
[+ant] [+cor] [−ant]
[−cor] [−del rel] [−del rel]

/t/ and /k/ require four features and /p/ requires five features in order to distinguish each from the others and from all of the other phonemes of English. As a second example, the segments /p, b, f, v, m, w/ in English are specified as [+, +labial], but any one of these segments will require one or more additional features to uniquely distinguish it from all the other segments.

Given this definition of a natural class, one should expect to find language-specific evidence to support the contention that two or more given segments constitute a natural class. While the sharing of a phonetic property, as ascertained in the phonetics laboratory, is in itself considerable evidence, one looks to find phonological corroboration of any phonetic relationship established by other means.

In general, we can say that two segments belong to a natural class when one or more of the following criteria are met in a number of languages:

1. the two segments undergo phonological rules together;
2. one or more of the following features are common to both segments in all the languages examined: [−voice, −cont, −del rel];
3. one or more of the following features are common to both segments in all the languages examined: [+labial, −cor];
4. the two segments share the same distribution of phonological features in a number of languages.

1. The feature [−del rel] is required to differentiate the stops /p, t, k/ from the affricate /ʧ/.
2. As mentioned in 2.5.1, the feature [+labial] can be substituted for [−cor, −ant], in which case /p/, /t/, and /k/ have an identical complexity of 4.
b the two segments function together in the environments of phonological rules;

c one segment is converted into the other segment by a phonological rule;

d one segment is derived in the environment of the other segment (as in cases of assimilation).

While these criteria are not foolproof, they more often than not serve as the basis for establishing natural classes.

Consider as an example the following phonological rule:

\[
\begin{align*}
{k} &\rightarrow [\check{c}] / -\{i\} \\
{g} &\rightarrow [\check{j}] / -\{e\}
\end{align*}
\]

Recalling the conventions discussed in 4.3.1.2.1 and 4.3.1.2.2, this rule is an abbreviation for the following four subrules:

\[
\begin{align*}
1 & \quad {k} \rightarrow [\check{c}] / -\{i\} \\
2 & \quad {k} \rightarrow [\check{j}] / -\{e\} \\
3 & \quad {g} \rightarrow [\check{j}] / -\{i\} \\
4 & \quad {g} \rightarrow [\check{j}] / -\{e\}
\end{align*}
\]

By criterion a, /k/ and /g/ constitute a natural class (the class of velar stops), since they undergo this phonological rule together. By criterion b, /i/ and /e/ constitute a natural class (the class of front unrounded vowels), since they function together in the environment of this rule. By criterion c, /k/, /g/, /\check{c}/, and /\check{j}/ constitute a natural class (the class of [+high] noncontinuants), since the first two are converted into the second two by this rule. And, finally, by criterion d, /\check{c}/, /\check{j}/, /i/, and /e/ constitute a natural class, since the first two are derived by this phonological rule in the environment of the second two. We can refer to this class as the class of palatals, although it should be noted that the Chomsky and Halle (1968) specification [+high, -back] excludes /e/ from the class.

It is important to note that these criteria are valid only if the rule in question is “natural” (see 5.2). That is, we must make sure that the rules upon which we base our supporting evidence for natural classes are frequent and plausible, and not unnatural or “crazy” rules. (For a discussion of how unnatural rules, that is, rules which are not phonetically plausible, come into a phonology, see 5.2.6).

Having established these criteria for natural classes, we can now examine the relationship between simplicity and naturalness. Since natural classes are formally defined by the feature-counting simplicity metric, it should generally be the case that classes which require fewer features to specify them are more natural than classes which require more features. While this sometimes turns out to be true, there are a number of cases where the simplicity metric breaks down.

Consider, for example, the following possible natural classes, arranged vertically:

<table>
<thead>
<tr>
<th>CLASS A</th>
<th>CLASS B</th>
<th>CLASS C</th>
<th>CLASS D</th>
</tr>
</thead>
<tbody>
<tr>
<td>b b b b</td>
<td>b b b b</td>
<td>b b b b</td>
<td>b b b b</td>
</tr>
<tr>
<td>d d d d</td>
<td>d d d d</td>
<td>d d d d</td>
<td>d d d d</td>
</tr>
<tr>
<td>g g g g</td>
<td>g g g g</td>
<td>g g g g</td>
<td>g g g g</td>
</tr>
<tr>
<td>[+voice]</td>
<td>[+voice]</td>
<td>[+voice]</td>
<td>[+voice]</td>
</tr>
<tr>
<td>[−cont]</td>
<td>[−cont]</td>
<td>[−cont]</td>
<td>[−cont]</td>
</tr>
<tr>
<td>[−nas]</td>
<td>[−nas]</td>
<td>[−nas]</td>
<td>[−nas]</td>
</tr>
<tr>
<td>[−son]</td>
<td>[−son]</td>
<td>[−son]</td>
<td>[−son]</td>
</tr>
<tr>
<td>[−syl]</td>
<td>[−syl]</td>
<td>[−syl]</td>
<td>[−syl]</td>
</tr>
<tr>
<td>[+voice]</td>
<td>[+voice]</td>
<td>[+voice]</td>
<td>[+voice]</td>
</tr>
</tbody>
</table>

As one goes from left to right, the natural class gets simpler (or more general). Thus, it takes three features to specify class A (the class of voiced stops), but only one feature to specify class D (the class of all voiced segments). Classes B and C are intermediate, each requiring two features. As can be seen from the list of segments, class C (the class of voiced nonvowels) is more general or inclusive than class B (the class of voiced obstruents), though this is not directly revealed by the simplicity count.

If feature counting suffices in itself as a criterion for evaluating the naturalness of a class, then class D should be the most natural and class A the least natural. We should then expect class D to satisfy the four criteria stated above more readily than classes A–C. However, upon close examination, it becomes increasingly difficult to find phonological rules referring to the classes A through D as one goes from left to right. That is, it is easiest to find rules which refer to class A and class B, but it is less easy to find rules which refer to class C, and almost impossible to find rules which refer to class D. The voiced stops of class A, for instance, are required as a class in order to state the process of intervocalic spirantization found in many languages. Its general form is as follows:

\[
\begin{align*}
3 & \quad [\check{b}] \rightarrow [\check{p}] / V \rightarrow V
\end{align*}
\]
The voiced obstruents of class B are required as a class to state the process of syllable-final devoicing in German and other languages:

\[
\begin{array}{c}
\text{f} \\
\text{t} \\
\text{k} \\
\text{g} \\
\text{d}
\end{array}
\rightarrow
\begin{array}{c}
\text{b} \\
\text{p} \\
\text{t} \\
\text{g} \\
\text{d}
\end{array}/-\$
\]

It is difficult to find a phonological rule which has as its input the class of voiced nonvowels (class C), while no language appears to require the class of all voiced segments (class D) in phonological rules.

Feature counting thus fails to provide an adequate hierarchy of natural classes. The most simple class (that is, requiring the fewest features) is the least natural (judging from the four criteria proposed above); similarly, the least simple class is the most natural (see Chen, 1973a:226).

Another indication of the weaknesses inherent in the feature-counting approach to natural classes is found in cases where opposite feature values define classes of differing degrees of naturalness. For example, the class of [+nasal] segments in many languages includes /m/, /n/, and /ŋ/. This class is considerably more natural than the class of [−nasal] segments, which includes non–nasal stops, fricatives, affricates, glides, liquids, and vowels. While the following commonly attested rule of vowel nasalization reveals that /m, n, ŋ/ constitute a natural class,

\[
V \rightarrow \ddot{V} /-\frac{m}{n}\frac{ŋ}{n}
\]

it is hard to imagine a phonological rule affecting all segments except /m, n, ŋ/. Similarly, the feature [+glottalic] may define a class of implosives in a language (for example, /β/ and /ð/). While these segments do constitute a natural class and are expected to function together in phonological rules, the class of [−glottalic] segments, that is, all segments except the implosives, is not natural. This asymmetry in the feature specifications characterizes most oppositions which were defined as privative by Trubetzkoy (see 2.2.2). That is, whenever a class of segments carries a “mark” which other segments do not carry, the “marked” class is a natural one, but the “unmarked” class is not as natural (and in fact can be quite unnatural). As will be seen below, the theory of markedness developed by Chomsky and Halle (1968) is an attempt to remedy some of the problems created by feature asymmetries.

### 5.1.2 Natural Segments

Since the evaluation of natural classes by feature counting failed to take account of the “intrinsic content” of the various feature specifications being evaluated, phonologists turned their attention next to natural segments.

As pointed out in Chapter 1, certain segments are more frequently attested in languages than others. Thus, the vowels /i/ and /u/ are more frequent (and hence more “natural”) than the vowels /ü/ and /au/. In general, a language will not have /ı/ or /u/ unless it already has /i/ and /u/. Similarly, it is hypothesized that children acquiring native Turkish (which has all four high vowels) will first learn /i/ and /u/ and only later the less natural vowels /ü/ and /au/. Historically, we expect these less natural segments to merge (context-free) with more natural segments; for example, /ı/ has become /i/ in Yiddish (compare German [Füüs], Yiddish [Fis] ‘feet”).

#### 5.1.2.1 Prague School Markedness

Since much of the discussion of markedness in recent works centers around the notion of “markedness,” it is helpful to trace the evolution of this concept. The original Prague School notion of “markedness” owes its existence to the phenomenon of neutralization discussed in 2.2.3 and 3.2.2. It is recalled that, in Trubetzkoy’s terminology, certain oppositions are constant while others are neutralizable. In addition, when two phonemes are neutralized in a given position, it is the “unmarked” member of the opposition which is found phonetically. Since German neutralizes /p, t, k, f/, and /b, d, g, v, z/ syllable-finally as [p, t, k, f, s], the voiceless obstruent series is said to be unmarked (in German). Since voiced and voiceless obstruents do not contrast in this position in German, Prague School phonologists would set up five archiphonemes (see 3.2.2) /P, T, K, F, S/, that is, phonological units which are unspecified for voice but otherwise contain all of the feature specifications shared by voiceless and voiceless obstruents. Frequently, it is the opposition member which “lacks” some phonetic property (in the sense of Trubetzkoy’s privative oppositions—see 2.2.2) which is found in the position of neutralization. Thus, Trubetzkoy (1939) distinguishes between an archiphoneme plus null (unmarked member of the opposition, for example, /t/ in German) and an archiphoneme plus a certain feature (marked member of the opposition, for example, /d/ in German).

In general, then, the unmarked member of an opposition is found in positions of neutralization. Translated into distinctive features, according to this view, the + value will usually be the marked value (since it indicates the presence of some phonetic property in privative oppositions), while the – value is the unmarked value (since it indicates the absence of some phonetic property in privative oppositions). However, this is not always the case. For example, some languages exhibit an opposition between oral and nasalized vowels only after oral consonants (for example, Nupe and some dialects of Chinese [Yen, 1968]). Thus, the following oppositions are found in Nupe (Hyman, 1972a:186):

\[
\begin{array}{ll}
[ba] & \text{`to cut'} \\
[bá] & \text{`to break'} \\
[má] & \text{`to give birth'}
\end{array}
\]

\[
\begin{array}{ll}
[da] & \text{`to get wet'} \\
[dá] & \text{`to be in'} \\
[ná] & \text{`to shine'}
\end{array}
\]
As seen in these examples, /a/ and /ã/ contrast after the oral stops /b/ and /d/ but not after the nasal stops /m/ and /n/. Instead, only nasalized vowels are found after nasal consonants. In other words, the vowels /a/ and /ã/ are neutralized after nasal consonants as [ã]. Must we therefore conclude that /ã/ is unmarked and /a/ marked?

What is important is that the expected member of an opposition should be viewed as unmarked *in a specific environment*. Thus /p, t, k, f, s/ are unmarked syllable-finally but may be marked intervocically, since many languages show a tendency to voice intervocalic consonants. The nasalization example in Nupe shows, however, that marked does not necessarily mean +, nor does unmarked mean −.

In the Prague School conception, markedness was a language-specific property. While later phonologists have emphasized the universality of markedness judgments (for example, /t/ is universally unmarked, /d/ universally marked), the evaluation of an opposition as one between marked and unmarked members depends crucially, in Prague School phonology, on the presence of neutralization. Trubetzkoy (1936:192) states this principle as follows:

I emphasize that unmarked and marked members of an opposition exist only in the case of neutralizable oppositions. Only in such cases does the distinction between unmarked and marked members of an opposition have an objective phonological existence. Only in this case is it possible to determine the feature of a phonological opposition with complete objectivity and without the assistance of extralinguistic means of investigation. If a phonological opposition is constant, the relationship between its members may sometimes be thought of as a relationship between unmarked and marked. However, this remains only a logical or psychological fact but is not a phonological fact. [translation by L. M. H.]

In this passage, Trubetzkoy's view of the phoneme as a *phonological* (rather than phonetic or psychological) reality becomes evident (see 3.2). In a language such as Nupe, which never neutralizes /p/ and /b/, there is no phonological reason to speak of /b/ as being marked. In English, on the other hand, since /p/ and /b/ are neutralized as [p] after /s/, for example, spin, this constitutes a phonological criterion for labelling /b/ as marked and /p/ as unmarked.

Phonetically, of course, /b/ carries voicing while /p/ lacks voicing. Also, speakers may "feel" that /b/ is marked, in that it is phonetically more complex. However, because of Trubetzkoy's position on phonological reality, the solution must be dictated by the sound system and not by universal phonetic or psychological criteria. According to him, where there is no language-specific evidence for setting up a markedness contrast, such an analysis is unwarranted. Martinet (1936:52), a disciple of Trubetzkoy, sums up essentially the same position: "Where the phonologist has not found any neutralization, he can of course indicate the existence of two parallel phonological series, but would be better off not to speak of markedness and archiphonemes." [translation by L. M. H.]

The assignment of markedness values is not always as straightforward as it may seem, however. Martinet (1936) argues that /t/ is marked in French and /d/ unmarked. He cites examples such as [mett] *medecin* 'doctor,' where he claims that the [t] is lax and unvoiced. Normally, /t/ and /d/ have the following feature specifications in French:

\[
\begin{array}{cc}
| /t/ | /d/ | \\
| --- | --- | \\
| [−voice] | [+voice] | \\
| [−tense] | [+tense] |
\end{array}
\]

Thus, from a logical point of view, /t/ could be unmarked (because it lacks voicing) or marked (because it is fortis, or [+tense]). Martinet argues for the second interpretation.

In addition to the above problem in analyzing markedness values, a further problem arises when there is neutralization in two different positions, and when the phonetic realizations in the two positions are not identical. Such an example is found in German. We have already seen that /s/ and /z/ neutralize in syllable-final position as [f]; /s/ and /z/ also neutralize in word-initial position in German, but this time as [f]. Just as no words end with [z] in German, no German words begin with [s]. In fact, it is only intervocically that /s/ and /z/ contrast, for example, *reissen* [r̩esən] 'to tear' vs *reisen* [raɪzn] 'to travel.' On the basis of the final neutralization, one might suggest /s/ as the unmarked member of the opposition, but on the basis of the initial neutralization, /z/ would be the unmarked member. In brief, then, in Prague School markedness, as in other approaches, there are indeterminate cases which do not fall neatly into place.

### 5.1.2.2 Universal Markedness

The notion of markedness developed by the Prague School has been elaborated and applied in a number of ways. To Praguians, markedness is defined in a language-specific way. Of course, it may be possible to look for universal tendencies in the way marked and unmarked values are assigned cross-linguistically, and in fact, such a study has been begun by Greenberg (1966b). On the other hand, the exact usage of the term "marked" has not been uniform.

At least four interpretations are assigned to the term "marked." The first view of markedness is that something which is marked is characterized by the addition of something, for example, /kʰ/ carries lip-rounding, while /k/ does not. In distinctive features it is [+round].

A second view of markedness is *frequency*. The unmarked member of an opposition occurs more frequently than the marked member. Thus Maddieson...
(1972:959) suggests that in a tone language, high tone is unmarked if it is more frequent than low tone; similarly, low tone is unmarked if it is more frequent than high tone. Proponents of this view of markedness will argue that /a/ is the unmarked vowel in a language where it has greater lexical (that is, in morphemes) and textual frequency than other vowels.

A third view of markedness is neutrality. In French, the epenthetic (inserted) vowel occurring nonetymologically as in *Arc de Triomphe [arkə da tri3f] is [ɔ]. Thus, schwa is the unmarked or zero (neutral) vowel in French, as opposed to [i] in Nupe (Hyman, 1970b) and [u] in Japanese (Lovins, 1973). As seen in the following examples (taken from Lovins, 1973:123),

<table>
<thead>
<tr>
<th>ENGLISH</th>
<th>JAPANESE</th>
</tr>
</thead>
<tbody>
<tr>
<td>paprika</td>
<td>papurika</td>
</tr>
<tr>
<td>public</td>
<td>burikku</td>
</tr>
<tr>
<td>pulse</td>
<td>parusu</td>
</tr>
</tbody>
</table>

the vowel [u] is generally inserted in Japanese when English words with unacceptable consonant sequences are borrowed. This also applies when the English word ends in a consonant, since Japanese permits only /n/ in final position.

A fourth view of markedness states that the unmarked member is the productive or regular one. In English, the unmarked (regular) pattern for di-syllabic nouns is to have stress on the first syllable (for example, *climax, sérpent*). In this fourth view, exceptions such as *ellipse* and *cémént* are marked with respect to stress.5

5.1.2.3 Markedness in Generative Phonology Starting with Chomsky and Halle (1968, Ch. 9) and Postal (1968, Ch. 8), markedness theory has come to play a central role in generative phonology (see also Cairns, 1969).

While generative markedness theory has its roots in Prague School phonology, there is at least one crucial departure: to generative phonologists, markedness values are universal and innate. Voiceless stops, as suggested by the universal of Jakobson (1941), are universally less marked than voiced stops, voiceless fricatives, etc. (see 5.1.2.1). Thus, markedness is no longer treated as a property of the phonologies of individual languages, but rather as part of general phonological theory, which aims to capture the linguistically significant generalizations characterizing sound systems. It derives its support from studies of universals in language acquisition, linguistic typologies, and linguistic change. Unmarked sounds are said to be generally acquired earlier than marked sounds by children. They are also generally required in the inventory of sounds of a language before marked sounds can be added. In linguistic change, sounds are seen as changing from marked to unmarked (for example, a context-free change from implosive */d* to [l]) or from unmarked to marked (for example, the context-sensitive change of */V* to [V] before nasal consonants).

In their epilogue, Chomsky and Halle (1968) propose that pluses and minuses be replaced by *u*'s (for unmarked) and *m*'s (for marked) in underlying representations. This theoretical reorientation is designed to resolve certain difficulties in the older approach. For instance, we saw in 5.1.1 that feature counting does not always lead to the establishing of clear natural classes. As a further example, compare the two natural classes given below, which are both statable using alpha notation (4.3.1.2.5):

\[
\begin{align*}
\text{aback} & \quad \text{around} \\
\text{low} & \quad \text{V}
\end{align*}
\]

\[
\begin{align*}
\text{aback} & \quad \text{ahigh} \\
\text{low} & \quad \text{V}
\end{align*}
\]

The first class of vowel segments is one which frequently needs to be specified, as was seen in the discussion of morpheme structure conditions in 4.2.1.2. The second class is highly unnatural and unexpected in languages. However, if the relative naturalness of these two classes is assessed in terms of the number of features required to specify them, we would have to conclude that the two classes are of equal naturalness. Since we know that this is not the case, the evaluation measure must be either revised or discarded.

To remedy this situation, Chomsky and Halle (1968) introduce marking conventions which are designed to evaluate the "intrinsic content" of the features. These conventions will judge [xback, around] as more highly
valued than [zback, zhigh], etc. Consider, for example, their marking conventions X and XI for vowels:

\[
X: [u \text{ back}] \rightarrow [+\text{back}] / [+\text{low}]
\]

\[
XI: [u \text{ round}] \rightarrow \begin{cases} 
[u \text{ back}] / [-\text{low}] & \text{a} \\
[-\text{round}] / [+\text{low}] & \text{b}
\end{cases}
\]

Convention X says that the unmarked (expected) value of the feature Back is [+back] if the vowel is [+low]. The reason for this is that the unmarked low vowel is /a/, which is [+back]. This vowel is more common and basic than the [-back] vowel /e/ or the [+round] vowel /a/ (which is also [+back], however).

Convention XI says that the unmarked value of the feature Round is (part a) identical with the feature specification for Back if the vowel is [-low], or (part b) [-round] if the vowel is [+low]. With the introduction of convention XI, the underlying specification for the vowel /i/ is now [-back, u round, -low], just as the vowel /u/ is now [+back, u round, -low]. In this new version of the theory, pluses and minuses cost one point each, as do m's. But u's are costless. Therefore, the above convention allows us to substitute a u for a + or - and thereby decrease the lexical complexity of items having the vowels /i, e, u, o/.

On the other hand, a vowel which is [-low] but which does not have the same feature value for Back and Round will be marked for the feature Round. We therefore have the following possibilities:

\[
\begin{array}{ccccccc}
\text{low} & \text{u} & \text{u} & \text{m} & \text{m} & \text{u} & \text{u} \\
\text{high} & \text{u} & \text{u} & \text{u} & \text{u} & \text{m} & \text{u} & \text{u} \\
\text{back} & \text{u} & - & \text{m} & \text{u} & - & \text{+} & \text{m} & - \\
\text{round} & \text{u} & \text{u} & \text{u} & \text{m} & \text{u} & \text{m} & \text{m} & \text{m}
\end{array}
\]

While vowels which are [-low] but not [zback, around] will automatically cost more than those which have the agreeing specifications for backness and roundness, there is no convention which assigns less cost to a [-low] vowel which is [zback, zhigh]. That is, there is no convention corresponding to Chomsky and Halle's convention XIa. Thus, this combination of alpha variables will automatically cost more than [zback, around], and the evaluation measure is thereby retrieved. Furthermore, as Chomsky and Halle (1968:403) note, since the marking conventions are universal and not part of an individual phonology, they are not assigned any cost, just as the brace and arrow notations are free and clear.

Turning to part b of convention XI, it is observed that the unmarked value of Round is [-round] if the vowel is [+low]. This results from the fact that /a/, which is [-round], is the unmarked low vowel. The vowel /e/ on the other hand, will be marked [m round], costing more than the [u round] vowel /a/.

This same approach is extended to a variety of features in 39 tentative marking conventions (Chomsky and Halle, 1968:404-407), for example, to the various places of articulation (where labial and dental articulations are less marked than velar articulations), to manners of articulation (where [ucont] is usually [-cont]), and to nasality ([u nasal] → [-nasal]).

### 5.1.3 Natural Systems

Having provided these marking conventions by which the u's and m's of underlying forms are converted into pluses and minuses, Chomsky and Halle turn to the naturalness of systems. Their concern is to account for the naturalness of vowel system (a) and the unnaturalness of vowel system (b):

\[
\begin{array}{cccccccc}
\text{a} & \text{i} & \text{u} & \text{e} & \text{o} & \text{e} & \text{o} & \text{a} \\
\text{b} & \text{u} & \text{u} & \text{m} & \text{u} & \text{u} & \text{u} & \text{e} & \text{e} & \text{a}
\end{array}
\]

The set of marking conventions they give for vowels defines the following matrix for the various vowel sounds examined:

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>i</th>
<th>u</th>
<th>e</th>
<th>o</th>
<th>e</th>
<th>o</th>
<th>u</th>
<th>a</th>
<th>e</th>
<th>o</th>
<th>u</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td></td>
<td>u</td>
<td>u</td>
<td>m</td>
<td>m</td>
<td>u</td>
<td>u</td>
<td>u</td>
<td>u</td>
<td>m</td>
<td>m</td>
<td>u</td>
<td>u</td>
</tr>
<tr>
<td>high</td>
<td></td>
<td>u</td>
<td>u</td>
<td>u</td>
<td>u</td>
<td>m</td>
<td>u</td>
<td>u</td>
<td>u</td>
<td>m</td>
<td>m</td>
<td>u</td>
<td>u</td>
</tr>
<tr>
<td>back</td>
<td></td>
<td>u</td>
<td>-</td>
<td>m</td>
<td>u</td>
<td>-</td>
<td>+</td>
<td>m</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>round</td>
<td></td>
<td>u</td>
<td>u</td>
<td>m</td>
<td>u</td>
<td>u</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
</tr>
</tbody>
</table>

From this table it is clear that /a/ is the unmarked vowel. This is well supported from acquisition studies, where /a/ is found to be the earliest acquired vowel, as reported by Jakobson (1941). Cross-linguistic typological studies of vowel systems also reveal that /a/ is apparently never lacking in any language.

The vowels /i/ and /u/ are considered to be only slightly marked, with a markedness value of 1 each, and the naturalness of these segments, as revealed by the marking conventions, accounts for the frequently attested...
triangular vowel system /i, u, a/ (Jakobson, 1941), which children construct early in their development. These vowels are also quite frequent and occur in nearly all languages. A number of vowels have a complexity of 2, and the last three vowels have a complexity of 3.

Chomsky and Halle (1968:409) propose the following principle to account for natural vowel systems: “The complexity of a system is equal to the sum of the marked features of its members.” Thus, returning to the two five-vowel systems, /i, e, u, o, a/ has a complexity of 6 (1 + 2 + 1 + 2 + 0), while /i/, /u/, /o/, /a/ has a complexity of 10 (2 + 3 + 2 + 3 + 0). The first vowel system is therefore seen to be more natural than the second.

However, there is a problem with this procedure, as Chomsky and Halle note, since the vowel system /i, u, a, / has a complexity of 6 (1 + 2 + 1 + 2 + 0), yet is not as natural as the vowel system /i, e, u, o, a/, which also has a complexity of 6. A second principle is therefore required.

The difference between the two systems is that the more natural system, after choosing the three least marked vowels /i, u, a/, chooses the vowels /e/ and /o/, which are marked with respect to height only. No judgment is made about whether [+back] or [-back] is less marked for nonlow vowels (see footnote 6). Thus /i/ and /e/ are entered simply as [-back] and /u/ and /o/ as [+back].

The second system also chooses the three least marked vowels /i, u, a/ and then chooses the vowels /i/ and /u/, which are marked not for height, as in the case of /e/ and /o/, but rather for roundness. In other words, what makes /e/ and /o/ more marked than /i/ and /u/ is that they are [-high], rather than [+high]; what makes /i/ and /u/ more marked than /i/ and /u/ is that /u/ is [+round] rather than [-round] and /i/ is [-round] rather than [+round].

While the two systems have an equal complexity, it is clear that the system with /e/ and /o/ is more natural and expected than the system with /i/ and /u/. Since counting marked features does not reveal this difference in naturalness, another principle is necessary, which Chomsky and Halle (1968:410) state as follows: “No vowel segment can be marked for the feature ‘round’ unless some vowel segment in the system is marked for the feature ‘high.’” This condition, as stated in absolute terms, rules out a vowel system /i, u, a, /; stated less absolutely, it correctly accounts for the relative unnaturalness of this system as compared to /i, e, u, o, a/.

Chomsky and Halle suggest that other such conditions may be needed. However, since there will be a number of principles, it is likely that the product of markedness feature counting, namely, the correct specification of /i, u, a/ as the unmarked three-vowel system, can also be captured by a principle. One could state that before any segments with an /m/ specification are chosen, the two vowels not having an /m/ specification (/i/ and /u/), which are marked respectively [-back] and [+back], must be chosen. In other words, feature counting is replaceable by other notions.

Of course, one problem is that it is not always clear which of two systems (for example, vowel systems) is more natural or highly valued. For example, which of the following two vowel systems is more expected?

\[
\begin{array}{cccccc}
\text{i} & \text{i} & \text{a} & \text{u} & \text{u} & \text{a} \\
\text{e} & \text{e} & \text{o} & \text{o} & \text{a} & \text{a} \\
\end{array}
\]

The system on the left has the unmarked three-vowel system, but also the corresponding three nasalized vowels /i, u, a/. The system on the right has the unmarked five-vowel system, but also the vowel /e/. In terms of markedness, the first system has a complexity of 5 (since /u nasl/ is [-nasl] for all segments), while the second system has a complexity of 8. Both are six-vowel systems, and yet it is not clear how nasalized vowels should be evaluated with respect to other relatively marked vowels. According to Chomsky and Halle’s conventions, /a/ has a complexity of 1, that is, [m nasl], while /i/ also has a complexity of 1. The same procedure of feature counting would lead one to conclude that the vowels /e/ and /o/, which have a complexity of 2, are more marked (that is, less natural) than /a/. This conclusion appears to be false, since the vowels /e/ and /o/ are more widely attested in languages than /a/. It is even more clear that /i/ is not equally marked with /a/, since /i/ is one of the vowels which is found in almost all languages, while /a/ is not found in most languages.

What this means is that [m nasl] represents more of a complexity than, say, [m high]. We are therefore faced with either assigning differential coefficients to the various features Nasal, High, etc., or seeking another condition or principle which would explain the greater complexity of certain nasalized vowels over certain oral vowels. It must be borne in mind, however, that some vowel systems may simply not be comparable, since their organizing principles are so different. Chomsky and Halle (1968) were careful to compare systems such as /i, e, u, o, a/ and /i, u, a, i, o, e/, where the parameters are relatively constant, that is, front/backness, height, and roundness. Introducing the parameter of nasality is not directly comparable, just as the introduction of retroflexion, pharyngealization, or tense/laxness may not be.

While certain segments are less natural than others, it is not likely that this observation will lead to a foolproof formula for evaluating the naturalness of systems.

The reason for this is that the complexity of a system is not a function of...
the complexity of the segments contained in it—at least not directly. Rather, natural classes and systems are natural because of the relationship between the segments. Consider the following matrix of the segments contained in it—at least not directly.

A number of observations can be made from this matrix. First, unmarked consonants are noncontinuant and unvoiced. That is, /p/ is considered to be less marked than either /b/, which is [m voice], or /f/, which is [m cont]. Second, anterior (labial and dental) consonants are less marked than nonanterior (palatal and velar) consonants. Thus, /p/ and /t/ are [u ant], while /k/ is [m ant]. Finally, no decision is made about whether the labial position is more or less marked for non-nasal stops than the dental position. Thus, as in the case of the front/backness distinction in nonlow vowels, the feature value for Coronal is entered as [−cor] for /p/ and /b/ and [+cor] for /t/ and /d/.

By the conventions Chomsky and Halle propose, there are five consonants which are marked for one feature only, namely /p, t, k, s, n/, about which they remark: “It is significant that these five consonants are rarely absent in the phonological system of a language” (1968:413).

This minimal consonantal system can of course be reinforced by consonants having a complexity of 2 each, as in the following two systems:

<table>
<thead>
<tr>
<th>p</th>
<th>t</th>
<th>k</th>
<th>b</th>
<th>d</th>
<th>g</th>
<th>f</th>
<th>s</th>
<th>x</th>
<th>m</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>ant</td>
<td>u</td>
<td>u</td>
<td>m</td>
<td>u</td>
<td>u</td>
<td>m</td>
<td>u</td>
<td>u</td>
<td>m</td>
<td>u</td>
</tr>
<tr>
<td>cor</td>
<td>−</td>
<td>+</td>
<td>u</td>
<td>+</td>
<td>u</td>
<td>m</td>
<td>u</td>
<td>m</td>
<td>u</td>
<td>u</td>
</tr>
<tr>
<td>cont</td>
<td>u</td>
<td>u</td>
<td>u</td>
<td>u</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>u</td>
<td>u</td>
</tr>
<tr>
<td>voice</td>
<td>u</td>
<td>u</td>
<td>u</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>u</td>
<td>u</td>
</tr>
<tr>
<td>nasal</td>
<td>u</td>
<td>u</td>
<td>u</td>
<td>u</td>
<td>s</td>
<td>m</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
</tbody>
</table>

Each of these two eight-consonant systems has a complexity of 11. Each also appears to be a natural system. While a has established a voice contrast, b has established a stop/fricative contrast as well as a nasal/oral contrast in two positions. The low figure of 11 reveals this naturalness. However, both c and d also add up to systems of a complexity of 11:

<table>
<thead>
<tr>
<th>e</th>
<th>p</th>
<th>t</th>
<th>k</th>
<th>d</th>
<th>p</th>
<th>t</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>d</td>
<td>g</td>
<td>s</td>
<td>x</td>
<td>f</td>
<td>s</td>
<td>m</td>
</tr>
<tr>
<td>a</td>
<td>n</td>
<td>m</td>
<td>u</td>
<td>n</td>
<td>m</td>
<td>u</td>
<td>m</td>
</tr>
</tbody>
</table>

Feature counting fails to reveal the naturalness of a and b as opposed to c and d. The latter two systems are unnatural because, unlike a and b, they are not organized according to principles of optimal contrast. The systems in c and d have arbitrarily incomplete series, for example, /s/ and /x/, but no /f/; /b/, but no /d/ or /g/. Thus, although c and d contain segments of equal naturalness to those in a and b, the resulting systems are not as natural.

Conversely, classes of segments can be equally natural even though they involve individual segments of greatly differing markedness values. The class of voiceless stops /p, t, k/ has a complexity of 3; the two classes /b, d, g/ and /f, s, x/ each have a complexity of 6, since voiced stops are marked for voice and voiceless fricatives are marked for continuance. The class of voiced fricatives /v, z, y/ has a complexity of 9, since these segments are [m voice] and [m cont]. However, each of these four classes is equally natural. A class (or system) is not defined by the complexity of the individual segments, but rather by the relationship between them. The segments /b/, /d/, and /g/, which are implosives, are highly marked and unnatural. Their occurrence in languages is considerably more restricted than that of any of the other classes discussed above—in fact, /g/ is very rarely attested (Greenberg, 1970).

However, if a language has these three implosive sounds, they constitute a class of equal naturalness to the class of voiceless stops.

### 5.2 Natural Rules

With the introduction of markedness theory into generative phonology, it became possible to formalize not only the naturalness of segments and systems, but also the naturalness of phonological rules, thereby distinguishing linguistically significant generalizations from spurious or nonsignificant ones. The first attempt to deal with natural rules was in the framework of linking conventions.

#### 5.2.1 Linking Conventions

Chomsky and Halle (1968:401) express the view that in the following examples the a rule is more natural as a phonological process than the b rule:

- **6a**: i → u
- **7a**: t → s
- **6b**: i → u
- **7b**: t → θ
However, as they point out, it is the rules in b which are simpler in terms of the number of distinctive features required to specify them:

\[ 6'a: [+high] \rightarrow [+back] \quad V \quad [+round] \]
\[ 6'b: [+high] \rightarrow [+back] \quad V \]
\[ 7'a: [+ant] \rightarrow [+cont] \quad C \quad [+strid] \]
\[ 7'b: [+ant] \quad +cor \quad C \quad [+strid] \]

Rules 6'a and 7'a each require one more feature than rules 6'b and 7'b. If Chomsky and Halle's judgments concerning the relative naturalness of these rules are correct, there appears to be a discrepancy between naturalness and simplicity. The more general rules in b, as judged by the fewest features, are not the more expected ones.

In order to remedy this inadequacy in the theory, Chomsky and Halle propose the notion of linking conventions. What makes 6a more natural than 6b is that the unmarked value of Round for nonlow vowels is identical with the specification of Back (see 5.1.2.3). What makes 7a more natural than is that the unmarked value of Strident is [+strident] for nonback fricatives and affricates (see below). Chomsky and Halle propose that rules 6a and 6b be rewritten as follows:\(^{10}\)

\[ 6'a: [+high] \rightarrow [+back] \quad V \quad [+round] \]
\[ 7'a: [+ant] \rightarrow [+cont] \quad C \quad [+strid] \]

In the reformulations in 6'a and 7'a, [+round] has been replaced by [u round] and [+strid] by [u strid]. This new formalism says that when a feature is changed (for example, [−back] to [+back] in 6'a), all other features which are dependent upon this feature change for markedness specifications can be changed to their unmarked value without adding any cost to the rule. Since [u round] is interpreted as [+round] when a nonlow vowel is [+back], 6'a “links up” with this marking convention and converts∕ to [u].

Similarly, 7'a links up with the marking convention for stridency, given below (Chomsky and Halle, 1968:407):

\[ XXVII \]
\[ [u strid] \rightarrow \left\{ \begin{array}{l}
[u \text{strid}] \\
[astrid] / [+ant] \quad [+cont] \quad +cor \end{array} \right\} \]

Part c of this convention states that stridency agrees with the specification for delayed release when the consonant is either [+ant] or [+cor], that is, when the primary point of articulation is prevelar (labial, dental, palatal). Affricates and fricatives are [+del rel] and so the unmarked value for Strident is [+strid] for these consonants. Since 7'a changes∕t∕ to a fricative, that is, [+cont] (and redundantly [+del rel]), it automatically “feeds” into marking convention XXVII. Thus, in the rule converting∕t∕ to [s], Strident can be entered as [u strid] and therefore not be counted by the simplicity metric.

Having reformulated rules 6'a and 7'a as the simpler rules 6'a and 7'a, the question now arises: how do we formulate rules 6'b and 7'b in this new framework? In order to show the complexity of these rules relative to rules 6'a and 7'a, it is necessary to somehow block the application of linking. Thus, the feature values [m round] and [m strident] are incorporated into the rules, as seen in 6'b and 7'b below:\(^{11}\)

\[ 6'b: [+high] \rightarrow [+back] \quad V \quad [m round] \]
\[ 7'b: [+ant] \quad +cor \quad C \quad [m strid] \]

Although there is no change in the feature specification of Round in 6'b, it is necessary to state the [m round] specification to the right of the arrow in order to correctly evaluate 6'b as costing 4 features, as opposed to the more natural 6'a, which costs 3 features (recall that [m F] costs one point, while

\(^{10}\) Chomsky and Halle actually leave [u round] unexpressed in the formalization of such a rule. When a [+high] vowel becomes [+back], the marking conventions automatically change the value of the feature Round to [−round] by means of linking. We shall incorporate [u round] into the rule formalism so as to avoid confusion with the pre-marking convention feature-saving formalisms discussed in 4.3.1.1.

\(^{11}\) Instead of [m round] and [m strident], which require an interpretation by means of the marking conventions, Chomsky and Halle use [−round] and [−strid], which will be automatically more costly than [u round] and [u strid] in rules 6'a and 7'a. Postal (1968: 184-185), however, discusses the use of [u F] and [m F] on the right of the arrow in phonological rules. We shall follow his proposal, since it creates less confusion with the earlier formalisms (see note 10).
156 Phonological Naturalness

5.2 Natural Assimilation Rules

An attempt to look directly at the process, rather than at the complexity of the output segments, is provided by Schachter (1969). Schachter points out that there are assimilatory processes which are natural and other assimilatory processes which are unnatural. He proposes that in the metatheory the n or "natural" value of a feature will be listed for any given feature in any relevant environment. For example, the two intervocalic spirantization rules 8 and 9 would be written (ignoring linking conventions) as follows:

\[ s' \rightarrow +voice / V - V \]
\[ t' \rightarrow +voice / V - V \]

The simplicity metric correctly reveals the first spirantization rule to be less complex than the second, since \[ s \] and \[ t \] are expected intervocalically. With linking convention XXVII applicable, however, the rules would have to be written as follows:

\[ s' \rightarrow +voice / V - V \]
\[ t' \rightarrow +voice / V - V \]

The linking convention applies to \[ t' \] to yield \[ v \] and \[ z \], while the first rule, which yields \[ s \] and \[ t \], includes the specification \[ m \] strident], and therefore costs more than the second rule. In other words, the linking approach provides the wrong relative naturalness judgment in this particular case. 13

Schachter's proposal is to replace the pluses and minuses derived by natural assimilation rules with the feature \( n \) (for "natural"). Thus, the two spirantization rules would be written by him as follows: 14

\[ s'' \rightarrow +voice / V - V \]
\[ t'' \rightarrow +voice / V - V \]

It is, of course, always possible that the marking convention should be revised, since Chomsky and Halle point out that their conventions are only tentative. Several of the major ones (not discussed here) have been revised by Cairns (1969), for instance.

12 The change of /d/ to [k] is, of course, parallel to the change of /t/ to [k] discussed in 7b above.

13 It is, of course, always possible that the marking convention should be revised, since Chomsky and Halle point out that their conventions are only tentative. Several of the major ones (not discussed here) have been revised by Cairns (1969), for instance.

14 That the \( [s] \) strident] specification may sometimes be needed in the resulting output of such a rule is seen from the common Bantu phenomenon by which \( [s] \) (which is \( +strid \)) is converted intervocalically to \( [y] \), not to the \( [s] \) fricative \( [f] \). In the few cases where the \( [s] \) fricative is found intervocalically, there is evidence for a secondary development by which \( [y] \) later changed to \( [z] \).
While both rules mention the features Continuant and Strident in their outputs, the \( n \), like the \( u \) for "unmarked," is not counted in assessing the complexity of rules. Thus, rule 8" costs 6 points (counting C and V as one each), while rule 9" costs 7 points. In order to phonetically convert \([n \text{ strid}]\) to a plus or minus, one looks at the list of \( n \) feature specifications and finds a statement to the effect that the natural value of Strident between vowels is \([-\text{strid}]\).

The real motivation for this proposal is Schachter's observation that many assimilatory processes are asymmetric in nature. One such example is palatalization. While it is a frequent phenomenon for velars to become fronted (palatalized) before front vowels, it is not a frequent phenomenon for palatals to become backed before back vowels. That is, rule 10 is natural, while rule 11 is unnatural:

10  \( k \rightarrow \epsilon / \{i, e, a\} \)
11  \( \epsilon \rightarrow k / \{u, o, a\} \)

Numerous languages convert /ki/ to /\( \hat{\epsilon} i \)/, while apparently no language converts /\( \hat{\epsilon} u \)/ to /\( k u \)/. When we attempt to formalize these rules in terms of distinctive features, two equally well-motivated rules involving the same number of features result:

10' \([+\text{back}] \rightarrow [-\text{back}] / [-\text{back}] \)
C  V

11' \([-\text{back}] \rightarrow [+\text{back}] / [+\text{back}] \)
C  V

Both rules involve 5 features, and yet rule 11' should be evaluated as infinitely more complex and unnatural than rule 10'.

With this problem in mind, Schachter (1969) proposes the introduction of the feature specification \( n \) (discussed above), which is evaluated as having no cost in feature counting:

10'' \([+\text{back}] \rightarrow [n \text{ back}] / [-\text{back}] \)
C  V

The revised rule in 10'' now states that a back consonant takes on the natural feature value of Back, in the environment "preceding a front vowel." Included in the necessary interpretive conventions will be one stating that the \( n \) value of Back is \([-\text{back}]\) before \([-\text{back}]\) vowels. The rule as now written then carries a complexity of 4 (since \( n \) does not count), while the rule written as 11' requires 5 features and is now formally more complex than the more expected rule. Notice that rule 11' cannot be rewritten as in 11' :

11'' \([-\text{back}] \rightarrow [n \text{ back}] / [+\text{back}] \)
C  V

### 5.2.3 The Relativity of Rule Naturalness

Studies into rule naturalness have revealed both the asymmetrical nature of assimilatory rules and the relativity of naturalness as a criterion for rules. The general form of an assimilation rule is seen in the following formula:

\[ X \rightarrow [nF] / [oF] \]

A segment X acquires the same feature specification as some feature F in the environment of a segment Y having that feature value. What this means is that X can become \([+F]\) in the environment of a Y which is \([+F]\), or can become \([-F]\) if Y is \([-F]\). However, many rules which are of this form are strikingly missing from phonological descriptions. One example we have already seen involves the failure of /\( \hat{\epsilon} i \)/ to become \([+\text{back}]\) (that is, [k]) before a \([+\text{back}]\) vowel. Another example of this type involves nasalization and denasalization.

One of the most natural rules in phonology is the nasalization of vowels before (or after) nasal consonants. As seen in 13,

13  \( V \rightarrow [+\text{nasal}] / [-\text{nasal}] \)
C

this rule is formally an assimilatory process, since a vowel acquires the feature specification \([+\text{nasal}]\) before a consonant which is \([-\text{nasal}]\). The denasalization of a consonant before an oral vowel is a less natural rule. While /\( \hat{\text{an}} \)/ quite naturally is realized phonetically as [\( \hat{\text{an}} \)], it is a rare occurrence to find /\( \hat{\text{an}} \)\!/ pronounced [da]. As seen in rule 14, however,

14  \( C \rightarrow [-\text{nasal}] / [-\text{nasal}] \)
V

this too is of the form of an assimilatory process. However, this and other such assimilatory processes are not found, or are rarely found, in languages.

In other cases there are assimilatory rules which are each natural but differ in their degree of naturalness (frequency, expectancy, etc.). Chen (1973a), Stampe (1972a), and Vennemann (1972d) have pointed out that palatalization...
and nasalization are highly dependent on vowel height. Consider, for example, the two palatalization rules 15 and 16:

15  $k \rightarrow \epsilon / - i$

16  $k \rightarrow \epsilon / - \{i\}$

In 4.2.2, it was argued that rule 16 is more simple (general) than rule 15. Thus, when these rules are formalized using distinctive features, rule 16' is judged to be more highly valued than rule 15' by the evaluation metric:

15' $[+\text{back}] \rightarrow [-\text{back}] / - [+$ \text{high} $- \text{back}]$

16' $[+\text{back}] \rightarrow [-\text{back}] / - [-\text{back}]$

Rule 15' requires 6 features, while rule 16' requires only 5. However, rule 15'/15' is clearly more natural than rule 16'/16', since it is found with greater frequency in the world's languages. As in the case of natural classes, a conflict is apparent between maximally simple and maximally natural rules. The more simple rule is less natural, and the more natural rule is less simple.

The palatalization case is particularly revealing of the factors at work in determining the naturalness and simplicity of a phonological process. Palatalization is more natural when it occurs only before /i/ because the vowel /i/ is more fronted than the vowels /e/ and /æ/. Since /i/ has the highest tongue position of front vowels, the process of palatalization will always take place before /i/. Many languages stop the palatalization process here, and this accounts for the great frequency of palatalization before /i/ only. On the other hand, some languages extend the palatalization to other front vowels, such that /k/ becomes [\epsilon] before /e/ (fairly frequently) and even conceivably before /æ/ (rare, but attested, for example, in French). The higher a front vowel, the more palatal it is, and the more likely it is to palatalize a preceding consonant. Thus, /i/ is more palatal than /e/, which is more palatal than /æ/. This hierarchy must be reflected in phonological theory if the current relative naturalness values are to be assigned to the various rules of palatalization.

A similar example revolves around the nasalization of vowels, as studied by Chen (1973a). Chen points out that, of the two rules 17 and 18,

17  $[+\text{back}] + [\text{nasal}] / - [\text{nasal}]$  (nasalization of /a/)

18  $V \rightarrow [\text{nasal}] / - [\text{nasal}]$  (nasalization of all vowels)

rule 17 is more natural, although 18 is simpler in terms of feature counting. In his investigation of Chinese dialects, Chen reports that some dialects nasalize only /a/ in this environment and that vowel nasalization typically begins with this low back vowel. In other words, before any other vowel can become nasalized before a nasal consonant, it is necessary for /a/ to nasalize. This view has been confirmed in the experimental work of J. Ohala (1971), who reports a greater propensity to nasalize among low vowels. Vowel nasalization thus appears comparable to consonant palatalization, except that the tendency to extend the nasalization process to all vowels is much greater than the tendency to extend the palatalization process to all front vowels. Most languages appear to extend nasalization to nonlow as well as low vowels, because of timing factors involved in the lowering of the velum. While the study of rule naturalness is in its infancy, it is clear that naturalness is not a binary property. Rules are more or less natural or more or less unnatural.

5.2.4 Strengthening and Weakening

Schane (1972) mentions, in addition to natural rules of assimilation, natural rules whose function is to preserve or create preferred syllable structures. It is often observed that consonants and vowels are subject to reduction in certain positions within a syllable or word, while they are relatively stable in other positions, often becoming reinforced phonetically. In order to capture such natural processes which affect syllabic and word structure, the traditional concepts of strengthening and weakening have been recently discussed within the framework of theoretical phonology (Foley, 1970; Vennemann, 1972a; Hooper, 1973). In particular, it has been suggested that different consonant types should be assigned strength values to capture “phonological relations” between segments, particularly (though not exclusively) as they function in syllables (see 6.1.1.1 for discussion of the syllable).

5.2.4.1 Preferred Syllable Structure As pointed out by Jakobson (1941), the unmarked syllable type is CV, that is, an initial consonant followed by one vowel. This is the only syllable type which is found in all languages; in addition, it is the first which is learned in child language acquisition, even in languages having other syllable types. Other syllable types are more or less marked or unnatural. A CVC syllable is somewhat unnatural, though it is frequently attested in languages. On the other hand, a VCCC syllable is considerably less natural and is found in relatively few languages.

Evidence for the relative naturalness of one syllable structure over another is seen from the kinds of phonological processes which are introduced in order to create or avoid various syllable types. Thus, rules of insertion or deletion of segments are natural to the extent that they produce more natural
The effect of this rule is to cause resyllabification just in case a Berber syllable would otherwise begin with two consonants; for example, *[agusnu] becomes *ag$gnu. We have already seen in the case of Yawelmani (see 4.4.2) that there is a derivational constraint against sequences of two consonants within the same syllable. Berber disallows sequences of two consonants at the beginning of syllables. The resulting syllabification VC$CV is seen to be more natural than $CCV.

The unnaturalness of CC sequences within syllables should, however, be qualified, since some sequences are tolerated much more than others. Thus, some languages permit syllable-initial two-consonant sequences, but only if the second consonant is a sonorant; for example, Ewe allows CLV and CGV syllables, while Gwari allows CNV syllables. In addition, the syllable type sCV is also attested in many languages. However, here there is good evidence that a language will tend to eliminate such a structure. Spanish, for instance, has a rule of vowel insertion of the following form:

\[ \emptyset \rightarrow s \mid # \rightarrow s C \]

The vowel \([e]\) is inserted before word-initial sC sequences, as seen in the following examples:

\[ /\text{spa}n/ \rightarrow [\text{espa}n] \quad \text{`Spain'} \]
\[ /\text{stu}f/ \rightarrow [\text{estufa}] \quad \text{`stove'} \]
\[ /\text{skwela}/ \rightarrow [\text{eskwela}] \quad \text{`school'} \]

Thus, instead of the unacceptable syllabification *[pga$na], we now have the acceptable syllabification [pga$na] (see Hooper, 1973:166-168). Rather than a syllable beginning with sC, we now have a syllable ending with [s], which conforms to the pre-existent syllable structure of Spanish (for example, [dos$] *`two*). Other languages which show a dislike for sC syllables include Hausa, Hindi, and Pidgin English. While Hausa has native words such as *fiska* [fis\$k\$] *face* with sC sequences separated by an intervening syllable boundary, syllable-initial sC sequences borrowed from English must be separated by an epenthetic vowel; for example, [s\$k\$\$] *school* (from English *scholar*). M. Ohala (1972:41) reports that the English loanword *station* appears in Hindi dialects as *[iste\$an], *[sa\$te\$an], or *[te\$an]. Finally, Pidgin English modifies Standard English *stick* as *[sitik], but *strong* as *[tron]. Thus, different strategies are utilized to avoid (break up) undesirable syllable structures.

While rules of insertion and deletion often serve the function of making syllable structure more natural, and are therefore said to be natural themselves, counter tendencies have been noted in the literature. The following rule of short-vowel deletion in Yawelmani was mentioned in 4.4.2:

\[ [\text{-long}] \rightarrow \emptyset / V C \_ CV \]

However, in converting a sequence VCVCV to VCCV, the syllable structure changes from [VSCVCV] to [VC$CV]. That is, three open syllables (which, recall, are favored by languages) are changed to a closed syllable followed by an open one. The relatively unnatural syllable structure VC is
obtained by this rule. In Grebo, the following alternations are found (Inees, 1966:3):

<table>
<thead>
<tr>
<th>SLOW SPEECH</th>
<th>RAPID SPEECH</th>
</tr>
</thead>
<tbody>
<tr>
<td>pédé</td>
<td>plé</td>
</tr>
<tr>
<td>bódó</td>
<td>bló</td>
</tr>
<tr>
<td>kidé</td>
<td>klé</td>
</tr>
<tr>
<td>gedá</td>
<td>gla</td>
</tr>
<tr>
<td>kpoda</td>
<td>kpla</td>
</tr>
<tr>
<td>gbudó</td>
<td>gbló</td>
</tr>
<tr>
<td>fodo</td>
<td>flo</td>
</tr>
</tbody>
</table>

"bald patch"        
"chalk"                
"chest"                
"divide"               
"sew"                   
"room"                   
"emptiness"

Grebo appears to be currently undergoing a vowel-deletion (syncope) rule by which a vowel is deleted in the environment $C - d V$ in rapid speech. As a secondary adjustment, resulting intermediate forms such as $pédé$ and $bdó$ are modified to $[plé]$ and $[bló]$. Thus, two CV syllables are now becoming one CLV syllable.

In both the Yawelmani and Grebo cases, vowel-deletion rules have led or are leading to less natural syllable structures, assuming that $CV$ is always the "preferred" syllable. These rules are not motivated by syllable structure considerations, but rather by considerations of word structure. In general, consonants are deleted and vowels inserted to facilitate natural syllabification. On the other hand, an unstressed vowel in a word can become reduced or deleted by a weakening process (see 5.2.4.3). Finally, a consonant can be inserted to separate two vowels (a hiatus) and thereby make two natural syllables out of a VV sequence. Thus, the following Spanish data from Hooper (1973:182)

<table>
<thead>
<tr>
<th>STANDARD (ASTORGA)</th>
<th>DIALECTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>veo</td>
<td>veyo</td>
</tr>
<tr>
<td>leo</td>
<td>leyo</td>
</tr>
<tr>
<td>creo</td>
<td>creyo</td>
</tr>
</tbody>
</table>

show the need for the following consonant insertion rule:

26 $\emptyset \rightarrow y / e \rightarrow o$

The result is two natural CV syllables.

5.2.4.2 Consonant Strengthening and Weakening  It has already been noted that a consonant is subject to strengthening and weakening processes relative to its position within syllables and words. Consider first the following changes frequently observed in intervocalic position:

27 $tappu > tapu > tabu > ta\breve{b}u > tawu > tau > to$:

16 The symbols $[e]$ and $[o]$ represent the so-called "muffled" vowels found in Grebo and related languages.

The processes illustrated in the above derivation are, respectively, (1) intervocalic degemination, (2) intervocalic voicing, (3) intervocalic spirantization, (4) intervocalic sonorization, and (5) intervocalic sonorant deletion. Finally, it is seen that a form such as $[tau]$ can further develop into $[to:]$ by vowel coalescence. The above processes are frequently referred to as intervocalic weakening. As the form progresses from left to right, the intervocalic consonant becomes more and more weak, until it finally drops out. Many of these types of weakening occur prevocally in Finnish, as seen in the following forms (Skousen, 1972a:571):

<table>
<thead>
<tr>
<th>STRONG FORM (NOMINATIVE)</th>
<th>WEAK FORM (GENITIVE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tapa 'custom'</td>
<td>tavan /tapa+n/</td>
</tr>
<tr>
<td>pato 'dam'</td>
<td>padon /pato+n/</td>
</tr>
<tr>
<td>sika 'pig'</td>
<td>sian /sika+n/</td>
</tr>
<tr>
<td>piippu 'pipe'</td>
<td>piipun /piippu+n/</td>
</tr>
<tr>
<td>lantti 'coin'</td>
<td>lantin /lantti+n/</td>
</tr>
<tr>
<td>kirkko 'church'</td>
<td>kirkon /kirkko+n/</td>
</tr>
</tbody>
</table>

As seen in the following informal rules,

28a $p \rightarrow v / \_ \_ \_ C$ $S$
| t \rightarrow d / \_ \_ \_ C$ $S$
| k \rightarrow $\emptyset / \_ \_ \_ C$ $S$

28b $pp \rightarrow p / \_ \_ \_ C$ $S$
| tt \rightarrow t / \_ \_ \_ C$ $S$
| kk \rightarrow k / \_ \_ \_ C$ $S$

prevocalic weakening takes place in Finnish if the following syllable is closed by a consonant (the genitive suffix $-n$ in the above examples). We observe, in 28a, intervocalic voicing of $[t]/$ to $[d]$, intervocalic voicing and spirantization of $[p]/$ to $[v]$ (a secondary development from $[\beta]$), and intervocalic loss of $[k]/$, through historical intermediate $[g]$ and $[y]$ stages. In 28b, geminates become degeminated intervocally when the following syllable is closed by a consonant.

On the basis of examples such as those above, we can propose the following definition of weakening: a segment X is said to be weaker than a segment Y if Y goes through an X stage on its way to zero.17 Strengthening, on the other hand, refers to the reinforcement of a segment, as when a nongeminate [p] becomes geminate or double [pp]. Skousen (1972a:569) reports the following strengthening rule (in Savo dialects of Finnish),

29 $C_1 \rightarrow C_2 / \breve{V} \_ \_ \_ VV$

which geminates a consonant following a short stressed vowel and followed

17 I owe this definition to Theo Vennemann.
by a long vowel or diphthong, for example, /têkəs/ 'he does' becomes [têkkəs].

As a second criterion for defining strong and weak segments, it is suggested that stronger segments or segment types are more resistant to weakening processes. On the basis of this observation and the definition given above, consonant types have been categorized according to strength scales based on place of articulation, manner of articulation, and states of the glottis. Foley (1970:90), for instance, provides the following matrix of strength values which are necessary "for a proper interpretation of the Germanic and Spanish consonant shifts":

<table>
<thead>
<tr>
<th>Phonological Strength</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k⁺</td>
<td>t⁺</td>
<td>p⁺</td>
<td>Y O B</td>
</tr>
</tbody>
</table>

The horizontal strength scale arranges consonants according to their place of articulation, while the vertical strength scale arranges consonants according to manner of articulation and voicelessness/voicedness. About vertical strength 4, Foley states: "The phonological elements k⁺, t⁺, p⁺ have diverse phonetic manifestation. They may appear as long stops kk, tt, pp (Italian, Finnish), as aspirates k⁺, t⁺, p⁺ (English), as affricates k', t', p' (German)."

Let us first consider the contention that labials are stronger than dental/ alveolars, which are in turn stronger than velars. Foley (pp. 88–89) cites evidence from Danish and Spanish showing that /b/ is stronger than /d/ or /g/. First, from Danish, Foley states that intervocalic voiceless velars and dentals become weakened, but not labials (at least not until recently): kage [kaye], English cake; bide [biːd], English bite; but kobe [kobe], English cheap. In the development of Spanish, intervocalic /g/ and /d/ have dropped, while /b/ remains (spirantized to [β]): Latin lego > leo 'I read,' credo > creo 'I believe,' but habere > haber 'to have.'

We have already seen, in our earlier discussion of intervocalic weakening, that geminates are stronger than nongeminate voiceless stops, which are stronger than voiced stops, which in turn are stronger than voiced fricatives, and which, finally, are stronger than voiced sonorants. Thus, Foley's vertical arrangement of consonant types seems motivated. Similarly, on the basis of the examples just seen from Danish and Spanish, the horizontal arrangement by place of articulation seems motivated for at least some languages. There are, however, two problems which should be singled out.

First, this hierarchy is in part language-specific. There is, in particular, good evidence that dentals are stronger than both labials and velars in some languages. Skousen (1972b:86), for instance, points out the following alternations between strong and weak consonants in the Adamawa dialect of Fula:

<table>
<thead>
<tr>
<th>Strong</th>
<th>Weak</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>f</td>
</tr>
<tr>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>k</td>
<td>h</td>
</tr>
<tr>
<td>b</td>
<td>r</td>
</tr>
<tr>
<td>d</td>
<td>Y / — [-back, V]</td>
</tr>
<tr>
<td>g</td>
<td>w / — [+back, V]</td>
</tr>
</tbody>
</table>

Of the six nonimplosive stops, only /t/ does not spirantize, suggesting that it is not only by its voicelessness stronger than /d/, but also by its dentality stronger than /p/ and /k/. Recall from the Finnish weak consonants [v, d, ʰ] that /k/ deletes entirely while /p/ voices and spirantizes. Since /t/ only undergoes one weakening process (voicing) in standard dialects, it can be argued that it is stronger here, too, that is, more resistant to loss. Interestingly, Vennemann's (1972a:6) proposed relative strength of consonants in Icelandic shows /t/ stronger than /p/ and /k/, as seen in the following scale:

<table>
<thead>
<tr>
<th>Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8</td>
</tr>
</tbody>
</table>

Perhaps it should not be surprising to see /p/ and /t/ play the strongest role in different languages, since it will be recalled that Chomsky and Halle (1968) did not distinguish either the labial or dental position as less marked than the other (see 5.1.3). However, in Luganda, the following situation obtains:

<table>
<thead>
<tr>
<th>Strong</th>
<th>Weak</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>β</td>
</tr>
<tr>
<td>d</td>
<td>1</td>
</tr>
<tr>
<td>J</td>
<td>y</td>
</tr>
<tr>
<td>g</td>
<td>g</td>
</tr>
</tbody>
</table>

Of the four places of articulation, /g/ is the only voiced oral noncontiguous not to weaken (for example, in intervocalic position, as in [olugάndá]). Since the velar position seems so much weaker than more front articulations (for example, Chen, 1973b shows that *m and *n merge with *y in final position), it is hard to predict the Luganda situation from such hierarchies. While Luganda may very well have once had a weakening of /g/ to [ɣ] and then lost it, the relative strength hierarchies fail to predict why the velar
position rather than the labial or dental position should revert back to the original stop articulation.

The second problem derives from the fact that relative strength values are assigned on the basis of weakening processes of a highly specific type. In particular, not enough attention is paid to the environment in which the weakening takes place. The following derivation represents the kind of weakening which takes place in word- or syllable-final position:

30 \( \text{ab} \to \text{ap} \to \text{at} \to \text{ak} \to \text{a?} \to \text{a} \)

The processes illustrated are (1) final devoicing, (2) final change from labial to dental articulation, (3) final change from dental to velar articulation, (4) final change from velar to glottal stop, and (5) loss of glottal stop. Similar processes of "final consonant depletion" are discussed by Maran (1971) for Burmese, while Chen (1973b) provides the following sketch of nasal and stop developments in final position in Chinese:

\[
\begin{array}{cccc}
\text{m} & \text{n} & \text{p} & \text{k} \\
\text{n} & \text{n} & \text{p} & \text{k} \\
\text{n} & \text{n} & \text{k} & \text{k} \\
\text{V} & \text{V} & \text{V} & \text{V} \\
\end{array}
\]

The labial and dental consonants first merge as dentals, which later become velar. Then final velars weaken to become a weak (almost nonexistent) nasal on the one hand or a glottal stop on the other. In the next stage, the weakened nasal and glottal stop fall, leaving respectively a nasalized vowel and a shortened vowel. In the last stage of the development, these vowels become denasalized and unshortened.

By far the most common final weakening process is devoicing. Since this process takes /b/ to [p] and then on to zero, /b/ must in this environment be interpreted as stronger than /p/, although in intervocalic position we have already seen it to be weaker. Thus, such strength scales are sensitive not only to the variations of individual languages but also to the exact environment in which the given segments occur, and strength scales such as those of Foley and Vennemann, if they are to have any use in phonology at all, must be made relative to a given position of weakness.

18 This is not to say that there are no cases of context-free weakening. One good candidate may be the following set of frequently attested changes:

\( \text{č} \to \text{s} \to \text{s} \to \text{h} \to \text{Ø} \)

19 Chen's model also allows for a merger of dentals and velars before the loss of the labials.

We have seen two positions in which weakening typically occurs: intervocically and word- and syllable-finally. In a CVCVC language, these two positions have in common that the consonant appears postvocically. In a language permitting consonant sequences, a consonant can become weakened postvocically or pre-pausally, for example, both consonants of a VCC syllable are potentially vulnerable to weakening. On the other hand, strengthening typically occurs word-initially and postconsonantally, that is, in positions where a consonant neither is preceded by a vowel nor occurs before a pause. A simple example from Korean will suffice.

The following strengthening rule of Korean

31 \[ \text{[−son]} \to \text{[+tense]} / \text{[−son]} \]

states that /p, t, k, s, č/ become tense or fortis \([p^+, t^+, k^+, s^+, č^+]\) when preceded by a noncontinuant obstruent, as in the following examples:

32 /saŋke/ → [sakt+a] 'to mix'
/čuŋke/ → [čukp+a] 'from the book'
/nopke/ → [nopk+e] 'highly'

This tensing of obstruents will, if anything, make \([p^+, t^+, k^+, s^+, č^+]\) less vulnerable to weakening and loss, and must therefore be seen as a strengthening process.

5.2.4.3 Vowel Strengthening and Weakening

We have seen a number of cases of insertion and deletion rules whose effect is to reduce consonant sequences within syllables. Schane (1972) mentions rules whose function is to maximize the perceptual distance between segments. A well-known case is the neutralization of certain vowels in unstressed syllables in some dialects of Russian, as follows:

\[ \text{[i]} \to \text{[e]} \to \text{[u]} \to \text{[o]} \to \text{[a]} \]  

As seen in this somewhat simplified account, there is a five-vowel contrast /i, e, u, o, a/ in stressed syllables, but only a three-vowel contrast /i, u, a/ in unstressed syllables. /e/ becomes [i], while /o/ becomes [a], when unstressed. While the fact that /e/ moves up in vowel height and /o/ moves down may seem to be asymmetric, the result is the unmarked three-vowel system /i, u, a/. These three vowels are generally claimed to be maximally distinct from each other (Jakobson 1941; see also Liljencrants and Lindblom, 1972). Since the absence of stress on a syllable tends to obscure the identity of the vowel (see 6.2.1.2.3), unstressed vowels may polarize around the most

20 I owe the following rule and examples to Kong-On Kim.
perceptually distinct vowel positions, that is, high-front-unrounded, high-back-rounded, and low-central-unrounded. Thus, rules which maximize the perceptual distance between segments are natural.

More generally, vowel reductions found in unstressed syllables are seen as weakening processes. In English, unstressed vowels tend to become lax and ultimately schwa, for example, away [swæ], conform [konform]. In the two pronunciations of the word cerebral, that is, [sérəbral] and [sərilə-brəl], notice that the unstressed vowels are pronounced with [ə]. In addition, in the history of English and French, final unstressed vowels weakened to schwa and then dropped. In French, final “e-muet” vowels are indicated orthographically, though usually not pronounced in the standard language; for example, petite [potit] ‘little (f.),’ fenêtre [fənɛːʁ] ‘window.’ Also, complex rules of schwa-deletion occur in the language which, for example, delete the first vowel of petite and fenêtre in the phrases la petite [la pətî] ‘the little one (f.),’ and la fenêtre [la fənɛːʁ] ‘the window.’ Reducing a vowel to schwa is therefore one step on the way to zero and is thus a weakening process.

Just as different consonants weaken more readily than others, different vowels are more or less vulnerable to reduction and loss. It should therefore be possible to give strength values to vowels as well as to consonants. Hooper (1973:170) proposes the following strength scale for Spanish vowels:

\[
\begin{array}{cccccc}
e & o & i & u & a \\
1 & 2 & 3 & 4 & 5
\end{array}
\]

As evidence for the relative strength of high vowels over mid vowels, Hooper cites the following weakening of high vowels in unstressed syllables:

33 LATIN SPANISH

díxi > dije ‘I said.’
lácus > lagos ‘lakes’
plicáre > llegár ‘to arrive’
lucráre > lográr ‘to succeed’

Since /a/ has never undergone reduction or deletion, Hooper concludes that it is the strongest vowel in Spanish. /e/ is said to be the weakest vowel, since it “has been deleted in word-final position after certain single consonants, while /o/ and /a/ remain.”

In general, the same remarks made about consonant strengthening and weakening apply to vowels. Thus, strength scales are both language-specific and environment-specific. There has been a considerable growth of interest in studying the processes by which segments become reinforced or reduced. Thus, the notions of coloring and bleaching developed by Miller (1972, 1973) and Stampe (1972b) are roughly comparable to the above notions of strengthening and weakening. For a critique of these notions, see J. Ohala (1974).

5.2.5 The Phonetic Basis of Natural Rules

There are probably other kinds of natural rules which do not fit neatly into one of the above categories. However, rules which linguists generally agree are natural all have in common the property of being phonetically motivated. While there may be cases where the phonetic explanation of a process is not known, in general the rules which are said to be natural can be attributed to either articulatory or acoustic assimilations or simplifications. Let us return to a few cases of assimilation as illustrations.

The rules in 34 and 35 are frequently cited as being natural:

34  \[k \rightarrow k'/ -i\]
35  \[v \rightarrow ñ / -N\]

While these processes of palatalization before /i/ and nasalization before /N/ are discussed as natural phonological processes, relatively little attention is given to the fact that 34 and 35 represent processes which are universally present in all languages (see, however, Stampe, 1969, 1972b). That is, a [k] will tend to be somewhat fronted before [i] and a vowel will tend to be somewhat nasalized before a nasal consonant. These universal tendencies are, as such, not part of the phonologies of individual languages, but rather belong to the realm of universal phonetics.

While the source of such assimilations is seen to be phonetic and universal, a given language may focus on one or more of these in such a way as to make them part of a language-specific phonology. Something which was automatic or “intrinsic” can thus become nonautomatic or “extrinsic” (Wang and Fillmore, 1961). In other words, something which is usually predictable from universal phonetics can become predictable only from a language-specific phonological point of view. This process of phonologization, whereby a phonetic process becomes phonological, can be seen from a comparison of the fronting of [k] in English and Luganda.

While the [k] of the English word key [ki] is somewhat fronted, it is questionable whether English has a specific phonological rule which is responsible for this. Since the fronting is slight, it seems preferable to attribute it to universal phonetic constraints on sequences of velar consonants followed by [i]. Luganda, on the other hand, pronounces [k] before [i] with a particularly noticeable palatal offglide. Thus, /ékikópó/ ‘cup’ is pronounced [ék'ikópó]. Luganda has phonologized a phonetic variation which is usually predictable on universal grounds. While one would expect a [k] to be somewhat fronted before [i], it is up to an individual language to further modify—or exaggerate—the fronting. This is precisely what Luganda has done in this instance.

Thus, the reason natural rules are the way they are is that they are deeply grounded in the universal phonetic properties of speech. In some languages,
a universal constraint on phonetic sequences, as just seen, is exaggerated—until it can no longer be predicted solely on the basis of universal phonetics, but rather requires a language-specific statement (rule) in the phonology.

A particularly clear example of this involves the following rule, which frequently occurs:

$$36\quad V \rightarrow [+\text{long}]/ - [+\text{voice}]$$

This phonological rule owes its existence to another phonetic universal which says that vowels are universally longer before voiced consonants than before voiceless obstruents. Numerous phonetic studies have verified this point (Chen, 1970; Lehiste, 1970; Mohr, 1971). In most languages, however, this process represents simply a low-level phonetic statement about the language—again, not a phonological rule that is language-specific, but rather a part of universal phonetics. However, some languages will phonologize this intrinsic property of vowels before voiced consonants by exaggerating the degree of lengthening to such a point that it can no longer be attributed to universal phonetics. English is such a language.

As shown by Chen (1970), the vowel-length difference in minimal pairs such as *bat:*bad exceeds the normal intrinsic variation found in other languages. The following table represents, for the languages studied by Chen, the ratio computed from the average length of a vowel before a voiceless consonant divided by the average length of a vowel before a voiced consonant (that is, $V_t/V_d$):

<table>
<thead>
<tr>
<th>Language</th>
<th>$V_t/V_d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>.61</td>
</tr>
<tr>
<td>Korean</td>
<td>.78</td>
</tr>
<tr>
<td>French</td>
<td>.87</td>
</tr>
<tr>
<td>Spanish</td>
<td>.86</td>
</tr>
<tr>
<td>Russian</td>
<td>.82</td>
</tr>
<tr>
<td>Norwegian</td>
<td>.82</td>
</tr>
</tbody>
</table>

In all six languages, it is observed that the length of a vowel found before a voiceless consonant ($V_t$) is less than the length of a vowel found before a voiced consonant ($V_d$). As computed by Chen, the closer the ratio approaches 1.0, the less the discrepancy in vowel length in the two positions. Thus, in the above figures, French shows the least difference in vowel length, while English shows the most. Furthermore, it is observed that English falls significantly below the .83 ratio obtained by averaging the remaining five languages. It thus appears that English has extended this vowel-length difference beyond the normal range predictable from the phonetics.

It is not quite clear at this time whether there is an absolute threshold or whether phonologization simply tends to enhance such a discrepancy. Notice, however, that there is an explanation for the exaggerated behavior of English vowel length. Since there is a tendency in English to devoice final voiced obstruents (such as in the word *bad*), the vowel-length discrepancy

has come to assume a phonological role, and perhaps ultimately a phonemic role. As has been shown by Denes (1955), the vowel-length difference in such pairs as *bat:*bad is much more important perceptually than any voicing difference which may be present in the final C. It is also relevant here to note that the initial contrast in the minimal pair *pat:*bat has been shown to be, perceptually, one of aspirated vs. unaspirated, rather than voiceless vs. voiced. It thus appears that English is in the process of losing its voice contrast in consonants (note the loss of the /t/-/d/ contrast in most intervocalic positions): the final voice contrast is being replaced with a length contrast and the initial contrast is being replaced with an aspiration contrast. Thus, in the terms of Jakobson (1931b), a "rephonologization" is currently taking place.

One way to show that a phonetic universal has been phonologized is to show that the language has exaggerated the degree of an intrinsic variation, as we have seen. Another way is to show that a further phonological process is in some way dependent upon the resulting phonetic distinction. The *bat:*bad example is quite appropriate. As shown by Labov, Yaeger and Steiner (1972), long vowels derived before voiced consonants (as well as long vowels derived in other contexts, for example, before /sf/) have become centralized diphthongs in certain Eastern dialects. Thus, while [bat] remains as such, [ba:d] (which tends to be pronounced [ba:t]) becomes [be:ed]. Since the phonological process of diphthongization must refer to the vowel-length distinction, this means that the lengthening of the vowel in *bad* must be part of English phonology.

### 5.2.6 The Denaturalization of Natural Rules

Such examples can be extended. The basic conclusion is that universal phonetic considerations usually provide the initial motivation for phonological rules, and since this is the case, there will be many rules written in synchronic phonologies which have this aspect of phonetic plausibility. While we have witnessed considerable interest in revealing and formally accounting for the fact that some phonological rules are phonetically plausible and others are not, there has recently been a critical reaction against the use of naturalness as a phonological criterion (Bach and Harms, 1972; Skousen, 1972a). One argument which is raised in this regard is the fact that natural rules tend to lose their naturalness through time. In this section, we shall discuss three mechanisms by which rules tend to become denaturalized: telescoping, morphologization, and rule inversion.

#### 5.2.6.1 Telescoping

The phenomenon of telescoping (Wang, 1968: 708) can be defined generally as the loss of an intermediate stage in a phonological derivation. A sound change occurs which interacts with a previously existing phonological rule so as to obscure the naturalness of the latter. An
example of this process can be seen in the palatalization of /k/ before [i]. While the original phonetically plausible rule is as follows,

37  \( k \rightarrow k^\prime / \_ i \)

the subsequent modification in 38 can enter into the phonology:

38  \( k^\prime \rightarrow \epsilon \)

The resulting phonological rule is the familiar one repeated in 39:

39  \( k \rightarrow \epsilon / \_ i \)

We have already seen that Luganda has rule 37, converting /ɛɛkÔ pó/ 'cup' to /ɛɛkÔ pó/. The tendency of rule 38 to further convert [k'] to [ɛ] is seen in the dialectal pronunciation of 'cup' as [ɛɛkÔ pó]. While the [i] environment has surely provided the motivation for the fronting of /k/, it has not provided the motivation for the affrication of [k'] to [ɛ]. Rather, two separate processes appear to be at work here. The first, as represented in 37, is a sequentially motivated rule, whereby /k/ assimilates in frontness to the following vowel. The second, represented in 38, is segmentally motivated, consisting of the context-free conversion of intermediate [k'] to [ɛ]. In the terminology of Vennemann (1972b), the first is an \( I \)-rule, since it increases the complexity of a segment, while the second is a \( D \)-rule, since it decreases the complexity of a segment. The segment [k'] is clearly more complex than the segment [k] (though not in the context /_i/). In addition, in the palatal position, affricates are more natural (or less "marked") than stops (Ladefoged, 1971:41). That it is not the vowel [i] that is directly responsible for the affrication is seen from the fact that [k'] is just as likely to become [ɛ] before [a] as it is to become [ɛ] before [i]. Thus Gwari speakers, whose language historically converted [k'wa] to [ca], frequently repeat the Hausa word [k'aw] 'beauty' as [caa].

The resulting rule in 39 is thus the result of the telescoping of the two processes in 37 and 38. In writing a rule such as 39, therefore, it should be recognized that an important historical intermediate form is by-passed. While the resulting rule in 39 still maintains a general phonetic plausibility (since phonologists speak of it usually as a palatalization process with only secondary affrication), telescoping sometimes leads to rules which are not only unnatural but simply "crazy" (Bach and Harms, 1972).

One such unnatural rule in certain Bantu languages concerns consonant changes which occur before Proto-Bantu */*. We shall limit our attention to the following subpart of a rule found in certain of these languages:

40  \( p \rightarrow s / \_ i \)

This rule states that /p/ is realized as [s] before /i/. While certain Bantu languages show an alternation between [p] and [s] and presumably therefore a need for rule 40, this rule is highly unnatural. While the rule of assimilation in 41 is attested in languages,

41  \( t \rightarrow s / \_ i \)

the change of a labial stop to an alveolar fricative is a relatively rare occurrence. Furthermore, if we wished to hypothesize the derivation in 42,

42  \( pi \rightarrow ti \rightarrow si \)

then the change of [p] to [t] before [i] is apparently unattested in languages. Rule 40 represents a telescoping of the following sound changes:

43  \( pi > p'i > p'i > t'i > si \)

The steps involved are (1) aspiration of obstruents before the high vowels /i/ and /u/, (2) affrication with an [s] release conditioned by the "grooved" vowel [i], (3) assimilation of place of articulation of the closure to the release of the affricate, and (4) deaffrication. Each of these historical changes is phonetically motivated, though the telescoped product in 40 is not natural in itself. However, since there is no reason to go through all the historical stages of 43 in a synchronic description, the form of the rule in 40 is adopted, however unnatural it may look on the surface.

5.2.6.2 Morphologization The example just discussed shows how a phonetically plausible rule (for example, the development of a "noisy" obstruent—such as an aspirated consonant—before a high vowel) can become less plausible, and eventually implausible. The resulting rule is, however, still statable as a phonological rule using only phonetic information and grammatical boundaries. A second way in which a phonetically plausible rule can become modified is for the environment to be reinterpreted as a morphological one, a process known as morphologization (Kiparsky, 1972; Robinson, 1972; Skousen, 1972a; Vennemann, 1972c, 1973; Hooper, 1973). The classic example is Umlaut in German, where the plural of Gast [gast] 'guest' is Gäste [gesta].

The derivation of the plural form [gesta] is seen in 44:

44  gasti > gesti > gesta 'guests'

Historically, the plural suffix on 'guests' was, phonetically, [i]. This [i],

11 Many Bantu languages obscure the original motivation for these changes by merging *i and *e as /i/. Thus, some instances of [i] will condition the consonant changes, while other instances will not.
after fronting \([a]\) to \([\varepsilon]\), was reduced to a schwa. Thus, the original rule was phonetically plausible, as in 45:\textsuperscript{22}

\[
45 \quad a \rightarrow \varepsilon / - C_t i
\]

The present-day rule must be stated in nonphonetic terms, as in 46:

\[
46 \quad a \rightarrow \varepsilon / - \text{morphological information such as [+pl], etc.}
\]

Since some nouns with \(/a/\) in the singular do not take an Umlaut in the plural, as seen in 47,

\[
47 \quad \text{With Umlaut} : \text{Gast/Gäste} \ '\text{guest/guests}'
\]

\[
\text{Without Umlaut} : \text{Tag/Tage} \ '\text{day/days}'
\]

it is even necessary to put a diacritic feature on \(\text{Gast}\) so that it will undergo the rule in 46.\textsuperscript{23}

Morphologization is a common phenomenon, and it is often a particular kind of telescoping. In the above example, the Umlaut rule and the schwa reduction rule have telescoped to yield the rule in 46.

\textbf{5.2.6.2 Rule Inversion} A rule which is morphologized is automatically “unnatural,” since it is not phonetically motivated. In addition, a rule which changes \(/a/\) to \([\varepsilon]\) in the plural certainly cannot frequently occur in the world’s languages. The third mechanism by which a natural rule can lose its phonetic plausibility is by \textit{rule inversion} (Vennemann, 1972c). Like telescoping, an inverted rule is not necessarily unnatural, as we shall now see in a case from Fe\?fe\?-Bamileke reported by Vennemann (1972c) and discussed in detail in Hyman (1972b).

Part of a general rule of intervocalic weakening in Proto-Bamileke is given in 48:

\[
48 \quad \text{Proto-Bamileke} \ d \rightarrow 1 / V - V
\]

Since East Bamileke dialects later introduced a rule deleting word-initial vowels, as in 49,

\[
49 \quad \text{East Bamileke} \ V \rightarrow \emptyset / ### -
\]

one of the several consonant alternations found in Fe\?fe\? involves that between \([I]\) and \([d]\) exemplified in 50:

\[
50 \quad \text{Fe\?fe\?} \ [\text{luu}] : [\text{nduu}] \ '\text{to beg}’ (perfective/imperfective)
\]

\textsuperscript{22} In rule 45, \(C_t\) signifies “at least one consonant.” This formalism specifies the lower and upper limits on a sequence of identical segments, e.g., \(C_n\) stands for “at least \(m\) instances of \(C\), but not more than \(n\),” such that \(C_0\) means either no consonants or one consonant, etc.

\textsuperscript{23} Alternatively, \(\text{Tag}\) could have a \textit{rule exception feature} which would make it exempt from Umlaut. For the treatment of exceptions in phonology, see Chomsky and Halle (1968:172-176), Kiparsky (1968a), Kisseberth (1970b), Schane (1973b).

This alternation can be captured by either of the following two rules:

\[
\begin{align*}
\text{51a} & \quad d \rightarrow 1 / ### - \text{(approx. historical rule)} \\
\text{51b} & \quad 1 \rightarrow d / n - \text{(inverted rule)}
\end{align*}
\]

Rule 51a, which takes \(/d/\) as underlying and derives \([I]\) word-initially, is closer to representing the original historical rule than is the inverted rule 51b, which takes \(/1/\) as basic and derives \([d]\) after \([n]\). In this particular case it is the historical rule (slightly modified) that has become unnatural, since there is no phonetic motivation for \(/d/\) to become \([I]\) word-initially. On the other hand, the inverted rule is natural, since a \([+\text{cont}]\) segment such as \(/1/\) can assimilate to the \([-\text{cont}]\) specification of a preceding homorganic nasal, thereby becoming \([d]\).

Arguments for the solution in 51b were presented by Vennemann (1972c). While the motivation for rule inversion is discussed in 5.2.8, the following example from Schuh (1972) suffices to show that rule inversion, if a valid phenomenon, leads to rule denaturalization.

The history of Hausa and related Chadic languages has been characterized by a number of syllable-final weakenings, which Schuh (1972:390-391) summarizes as follows:

\[
52 \quad *P > w / - S \\
*T > r / - S \\
*K > w / - S
\]

Velars and labials in the reconstructed proto-language become \([w]\) syllable-finally, while proto-alveolars become a trilled \([R]\). These changes, known as Klingeneheen’s Law, are responsible for the following alternations:

\[\begin{array}{lll}
\text{SINGULAR} & \text{PLURAL} & \text{GLOSS} \\
\text{juujii} & \text{jibaajee} & \text{‘rubbish heap’} \\
\text{samaarrii} & \text{samaarrii} & \text{‘young man’}
\end{array}\]

The etymological labial consonant \(/b/\) or \(/m/\) is found in the plural form, while the syllable-final reflex \([w]\) is found in the singular form (where the \([u]\) of ‘rubbish heap’ can be analyzed as coming from intermediate \([iw]\)). If ‘rubbish heap’ and ‘young man’ are set up with the underlying forms \(/jibjii/\) and \(/samayii/\), then the singular forms can be predicted by the following rule:

\[
53 \quad [+\text{labial}] \rightarrow w / - S
\]

\textsuperscript{C}

This rule is phonetically plausible, since syllable-final weakening is a widespread process in languages. On the other hand, if we were to start with the underlying forms \(/jiwjjii/\) and \(/sawayiyi/\), we would run into two problems. First, the rule required to derive the plural forms would have to be stated in
such a way that /w/ sometimes would become [b], sometimes [m], that is, /jijwji/ would have to be marked with a diacritic [+B] and /sawrayji/ with a diacritic [+M], or else the incorrect forms *jimaaje and *sabaarit might result from the rule. Second, the resulting inverted rule would be exceedingly difficult to state and would require considerable morphological information. In the two plural forms given, it looks as though /w/ is becoming [b] and [m] intervocally. Since there is no phonetic reason for /w/ to become a stop in this position, the resulting inverted rule would therefore be unnatural. In conclusion, rule inversion can lead to either unnatural phonetic motivation of a phonological rule. The question might be raised at this point, why do languages permit the phonetic naturalness of processes to be destroyed? In other words, why don't they fight back?

The problem revolves around the question of whether naturalness is a valid synchronic criterion for evaluating phonological systems. Given the two hypothetical languages L₁ and L₂,

<table>
<thead>
<tr>
<th>L₁</th>
<th>L₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>sg. pl.</td>
<td>sg. pl.</td>
</tr>
<tr>
<td>ba a̱a</td>
<td>ąa aha</td>
</tr>
<tr>
<td>da a̱a</td>
<td>ąa ada</td>
</tr>
<tr>
<td>ga a̱a</td>
<td>ąa aga</td>
</tr>
</tbody>
</table>

no one would deny the naturalness of L₁ and the unnaturalness of L₂. The rules required are given below:

54 L₁ :  

<table>
<thead>
<tr>
<th>d</th>
<th>[b]</th>
</tr>
</thead>
<tbody>
<tr>
<td>g</td>
<td>[v]</td>
</tr>
</tbody>
</table>

55 L₂ :  

<table>
<thead>
<tr>
<th>[b]</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>g</td>
<td>[v]</td>
</tr>
</tbody>
</table>

It is more natural for stops to spirantize intervocally, as in 54, than it is for spirants to become stops intervocally, as in 55. The situation in L₁ is found much more frequently than the situation in L₂—which, in fact, may never be found.

The question is, is L₂ never found because of the intrinsic unlearnability of this synchronic (static) state or because there is no straightforward diachronic (historical) source for such a synchronic state? The natural system in L₁ is obtained from a single natural sound change of intervocalic spirantization. While it is possible to imagine a chain of phonetically plausible events which would give rise historically to L₂, as seen below,

56 sg. *aba > a̱a > βa (intervocalic spirantization, loss of initial vowel)

pl. *amba > aba (loss of nasal mark of plurality)

It is significant that L₂ would require the convergence of several changes. It may, then, be less frequently attested because it requires such a complex historical source.

In order to refute this statement, it is necessary to find evidence that naturalness does play a role in synchronic phonology. There appears to be only counter-evidence. First, we should take note of recent arguments to the effect that speakers often do not "capture" phonological relationships in terms of natural phonological rules. Skousen (1972a) presents cases in Finnish where speakers appear to prefer morphologized rules to phonetically plausible ones (see Kiparsky, 1973 for a critique of Skousen). Recall the weakening process in Finnish, which applies when an obstruct is in intervocalic position and followed by a closed syllable:

57    

The following forms are repeated from 5.2.4.2:

<table>
<thead>
<tr>
<th>nomina tive</th>
<th>genitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>tapa</td>
<td>tavan</td>
</tr>
<tr>
<td>pato</td>
<td>padon</td>
</tr>
<tr>
<td>sika</td>
<td>sian</td>
</tr>
</tbody>
</table>

The n-suffix of the genitive construction closes the second syllable and conditions the change from /p, t, k/ to [v, d, Θ].

As Skousen points out, all of the phonetic information included in 57 is available to speakers of Finnish—and yet, he claims, speakers seem not to have knowledge of the phonetic conditioning of the rule. Instead, they reinterpret the rule as applying in the presence of certain grammatical suffixes, among which are the genitive -n and the inessive -ssä, both of which close preceding syllables. As stated by Skousen (1972a:571), speakers "memorize that the genitive suffix n and the inessive suffix ssä take the weak form of the stem without even perceiving that both suffixes close the syllable."

One of Skousen's arguments derives from the observation that some dialects of Finnish degenerate the -ssä suffix to -sä but still maintain the weak consonants. The standard inessive form of /käte/ 'hand' is [kädesssä], which is syllabified as [kä̱desssä]. The dialectal form, where the weakened form of /t/ is [r] instead of [d], is [karesä], syllabified as [käṟreßä]. If
speakers had knowledge of the role of the closed syllable in the weakening rule, they would have changed the [r] of this form back to [t], that is, *[kätésə]. Although the second syllable is no longer closed, speakers do not show even a slight tendency to change [r] to [t]. Thus, Skousen argues that the weakening rule should be revised to include a grammatical environment:

\[
\begin{align*}
57' \quad [p] & \rightarrow [v] \\
[t] & \rightarrow [d] / \rightarrow V \text{genitive/inclusive/etc.}
\end{align*}
\]

It is claimed that in the standard dialect, where -ssā is maintained, this reanalysis has already taken place. Thus, when -ssā is modified dialectally to -sā, there is no rule change, only a phonetic change.

While Skousen’s examples mostly involve morphologization as the means of capturing a phonological alternation, a more extreme case from Ngwe-Bamileke (Dunstan, 1966) suggests that speakers were not aware of an alternation at all. Ngwe, as well as other West Bamileke languages, is characterized by a rule which deletes the schwa of the class 6 mā- plural prefix when the stem begins with a labial consonant:

\[
58 \quad mā- \rightarrow m / \rightarrow [+\text{labial}]
\]

Thus, in the following forms from the related Mbiu dialect,

\[
\begin{align*}
59 \quad /mā-\text{bù}[:] & \rightarrow [\text{rhbù}:] \quad \text{‘breasts’} \\
/mā-\text{sò}[:] & \rightarrow [\text{māsò}]:] \quad \text{‘teeth’}
\end{align*}
\]

the phonetic form of ‘breasts’ has lost the schwa of the /mā/ prefix (since the stem begins with a labial consonant), while the schwa remains in the form for ‘teeth.’ Ngwe dialect has modified this earlier state of affairs by replacing the mā- prefix by the more productive bò- prefix of another plural class. However, as seen in the following forms,

\[
\begin{align*}
60 \quad /mā-\text{sò}[:] & > [bāsò]:] \quad \text{‘teeth’} \\
/mā-\text{bù}[:] & > [\text{rhbù}:] \quad \text{‘breasts’}
\end{align*}
\]

it is only phonetic [mā] which is replaced by [bā], and not underlying /mā/. Since [rhbù:] does not become *[bābù:], it appears that speakers failed to see the relatedness of the syllable m– prefix of ‘breasts’ to the mā– prefix of ‘teeth.’ Thus, when morphological categories such as noun classes undergo levelling, this process takes place on the basis of the phonetic shapes of the prefixes and not on the basis of the underlying morphological identity of the prefixes. Stated differently, the reality of rule 58, which converts /mā/ to [m] before labial stems, is not revealed in language change.

The basis of Skousen’s argumentation can be recapitulated as follows. It is proposed that a given language has a rule of the following form:

\[
61 \quad A \rightarrow B / C
\]

The conversion of A to B is conditioned by an environment C. Now, if this environment is destroyed, that is, either modified phonetically or perhaps totally lost, there are two possibilities. First, the language can continue to derive B in the new environment, say D. Or, second, the language can change B back to A, since the conditioning environment is no longer present.24 Let us say, for example, that a language has the following rule of palatalization:

\[
62 \quad k \rightarrow ċ / \rightarrow i
\]

A form such as /paki/ will be pronounced [pacin]. Now, if the language undergoes a sound change converting \*[i] to [a] word-finally, the language can react in one of two ways. First it can convert [pacin] to [pacin], in which case palatalization appears, on the surface, to take place before [a]. Or, second, it can change [pacin] to [pakan]. In the second case, [č] has reverted back to [k], since the conditioning environment \*[i] is no longer present. This second alternative provides evidence for the psychological reality of the palatalization rule. When \*[i] started to change to [a], speakers, well aware of the fact that /k/ was converted to [č] because of this \*[i], undid this rule when \*[i] was no longer heard phonetically. If it can be shown that speakers undo rules in just such cases, then evidence is obtained for the reality of such natural rules. It should be noted, however, that the more frequent phenomenon is for the language to leave the derived segment in the new environment, that is, [paca].

A second type of evidence which might be sought for naturalness as a phonological criterion can be outlined as follows. Let us say that a language has the same rule of palatalization of /k/ to [č] before /i/ as in 62, and that the same change of final \*[i] to schwa is about to occur. If naturalness is a valid phonological criterion, it should be possible for a sound change to be inhibited only in environments where its application would destroy the phonetic naturalness of a previously existing rule. As seen in the following hypothetical forms,

\[
63 \quad /papi/ \rightarrow [\text{papi}] > [\text{papa}] \\
/paki/ \rightarrow [\text{pači}] > [\text{pači}] \quad \text{(not *[paca])}
\]

the [i] of [papi] becomes a schwa, but the [i] of [pači] does not, since the naturalness of the palatalization rule would be destroyed (that is, [č] would in the form *[paca] be derived before the vowel [a]). Although sound changes

24 A third possibility which should be mentioned is that there will no longer be a rule at all, i.e., A and B will become contrastive or phonemic.
are sometimes blocked by considerations within a paradigm (for example, so that singular and plural forms do not merge; see Vennemann, 1968b; Kiparsky, 1972:196–206), no corresponding force has been discovered which would strive to keep rules natural. Instead, the above examples show the great tendency for rules to become unnatural (see 5.2.6), that is, to lose their phonetic plausibility and become morphologically conditioned.

5.2.8 Rule Simplicity as a Phonological Criterion

Having questioned the validity of rule naturalness as a phonological criterion (that is, a criterion for what is more readily learned as a phonological rule), it is appropriate to return to the notion of phonological simplicity discussed in Chapter 4. In 5.2.6.2 and 5.2.6.3, the phenomena of morphologization and rule inversion were introduced. The question now arises, when are rules to be represented as morphologically conditioned or as inverted rather than as phonetically conditioned?

Unfortunately, the criteria for choosing between solutions are not entirely clear, although certain clues can be isolated. For instance, consider the Turkish data discussed by Zimmer (1970:91ff):

64 /söyIE + ýor/ → [söylüyö] ‘he is saying’
not *[söylöyö]

In the underlying form, /E/ stands for the archiphoneme “unrounded non-high vowel” (that is, /e/ or /a/), while /I/ stands for the archiphoneme “high vowel” (that is, /i/, /ui, /l/ or /u/). Although the starred form *[söylöyö] is predicted from the general rules of Turkish phonology, the form [söylüyö] is found instead. To account for this fact, Zimmer considers the possibility of an additional rule of the following form:

65 i → ü / [+round] — — [+round]
V V

The vowel [i] is rounded to [ü] when it appears between two rounded vowels. However, it turns out that the ýor suffix is the only grammatical context which will ever satisfy rule 65.25 Thus it is just as easy to represent this rule as morphologically conditioned:

65’ i → ü / [+round] ýor [ —
V

We have seen, in the Finnish example, that speakers may attribute an alternation to grammatical rather than phonetic environments; a phonetic environment satisfied by only one suffix is even more likely to be “mis-interpreted” by speakers, as in the Turkish example. In brief, then, when given the chance to capture a phonological alternation by either a phonetically or a grammatically conditioned rule, there is a tendency toward the latter (see Hyman and Schuh, 1974:94).

In his study of rule inversion, Vennemann (1972c) states that the major factor contributing to this reanalysis is semantic. Semantically “basic” categories tend to be construed as providing the base forms for phonological representations. To reveal the problem facing generative phonology, consider the two hypothetical dialects D1 and D2:

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>D2</th>
</tr>
</thead>
<tbody>
<tr>
<td>sg.</td>
<td>pap pak pa pa</td>
<td>pl. papi pači papi pakı</td>
</tr>
<tr>
<td>pl.</td>
<td>papi pači papi pakı</td>
<td>pl. papi pači papi pakı</td>
</tr>
</tbody>
</table>

In a standard account, we might set up the underlying forms /pap/, /pap+i/, /pak/, and /pak+i/ for both dialects. D1 would require the phonological rule in 66, while D2 would require the phonological rule in 67:

66 k → c / — i (palatalization)
67 C → O / — ## (final consonant deletion)

By 66, underlying /pak+i/ becomes [pači] in D1, and by 67, underlying /pap/ and /pak/ both become [pa] in D2.

Although each dialect starts with the same underlying forms and contains a phonetically-based phonological rule, the synchronic state represented in D2 is radically more complex than that in D1. In particular, since the underlying final consonants can be discovered only by knowledge of the plural forms, which are morphologically and semantically more marked than the singular forms, D2 poses a problem for the language learner which D1 does not pose. In D1, all the necessary information for the application of rule 66 is contained in the unmarked singular form /pak/. The plural pronunciation [pači] is therefore in a crucial way derived from the singular [pak] plus an [i] suffix. In D2, on the other hand, the two forms [pa] are derived from underlying representations which are based on the plural forms.26

If simplicity is to be maintained as a synchronic criterion, and if the notion of simplicity is designed to express the intrinsic difficulty or learnability of a language, then the standard account of D2 fails miserably. For

25 While Barbara Robson and Alan Harris have privately expressed reservations about this analysis to me, the problem of what Zimmer calls “accidental reference” to a single morpheme, which results when only one morpheme satisfies the conditions for a phonological rule, is an interesting one.

26 While the [pa]-[pa] homonymity is a problem, it is important to note that we are not talking about this complexity here. Rather, the problem under consideration is the impossibility of predicting the plural forms [papı] and [pakı] from the singular form [pa]. To better understand the fact that there are two distinct problems to differentiate, consider another language where the two singular forms [pak] and [pač] are both [pač] in the plural. This language would also have the homonymity problem, but this time in the plural form only. It would still, however, be possible to predict the plural from the singular forms. Thus, the problem which Vennemann claims leads to rule inversion is not found in this language.
there is nothing in the analysis of D₂ that suggests that it is any more complex than D₁. In the framework of rule inversion, rule 67 would be rewritten as 67’:

67’  \( \emptyset \rightarrow p, k, \text{etc.} / V \rightarrow V \)

The choice of the exact consonant would be dependent on diacritic features, such as /pa/₂, /pa/₃. Because of the intrinsic complexity of arbitrary morphological classes, this solution would reflect the less simple phonological system of D₂ as compared to D₁. While all of the discussion of Chapter 4 highlighted the attempts of linguists to reveal the simple and general properties of languages, it is important that a theory of language also reveal complex and nongeneral properties when they exist.

As a concluding example, let us return to the Maori data discussed in 3.4.1, which are repeated below:

<table>
<thead>
<tr>
<th>VERB</th>
<th>PASSIVE</th>
<th>GERUND</th>
<th>GLOSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>hopu</td>
<td>hopukia</td>
<td>hopukaŋa</td>
<td>‘to catch’</td>
</tr>
<tr>
<td>aru</td>
<td>arumia</td>
<td>arumaŋa</td>
<td>‘to follow’</td>
</tr>
<tr>
<td>tohu</td>
<td>tohuŋia</td>
<td>tohuŋa</td>
<td>‘to point out’</td>
</tr>
<tr>
<td>maatu</td>
<td>maaturia</td>
<td>maaturanga</td>
<td>‘to know’</td>
</tr>
</tbody>
</table>

As seen in the verb stem, all of the forms end in /u/. Yet in the passive and gerund the consonants [k, m, n, r] appear. In 3.4.1 it was argued that in order to predict these consonants the verbs should be represented in their underlying forms as follows: /hopuk/, /arum/, /tohuŋ/, and /maatur/. The following rule of final consonant deletion was proposed:

68  \( C \rightarrow \emptyset / -## \)

This is the solution which the standard model of generative phonology would lead one to assume. However, Kiparsky (1971), basing himself on Hale (1971), presents a number of indications that speakers are not storing lying forms with final consonants, but rather setting up distinct classes of suffixes, for example, /kia, mia/. In other words, it is argued that there is not a derivational relationship between forms such as [hopu] and [hopuk], which are found respectively word-finally and before a vowel, but rather a single form /hopu/, with the /k/ being assigned to the suffix rather than to the stem.

Following the principle of rule inversion, one could propose a rule of consonant epenthesis, whereby a consonant is inserted before certain suffixes (for example, /ia/ ‘passive’ and /aga/ ‘gerund’):

69  \( \emptyset \rightarrow \{ k, m \} / \text{passive} \{ n, r \} / \text{gerund} \)

Just as in hypothetical D₂, it would be necessary to place a diacritic on verbs stating which consonant they take. In fact, some verbs would have no diacritic, since they do not take a consonant, for example, /patu/ ‘to strike, kill’ has the passive form [patua], where the /i/ of the passive has dropped.

The solution representing the passive (and gerund) as a large number of suffixes (/kia, /mia, etc.) receives considerable support from the fact that only the /tia/ suffix (which is not illustrated above) is productive. Kiparsky (1971) gives six ways in which this productivity manifests itself:

1. Stems which are basically nominal are often used verbally in spontaneous discourse; when they are so used, in the passive, they regularly take the ending /−tia/. (2) Derived causatives (formed with the prefix /whaka−/) take /−tia/ in the passive even if the basic verb stem takes another alternant when not in the causative. (3) There is a rule whereby certain adverbials are made to agree in voice with the verbs they modify; these adverbials take /−tia/ in the passive regardless of the shape of the passive ending which the verb itself takes. (4) Borrowings from English, including unassimilated consonant-final ones, take the ending /−tia/ in the passive. (5) Compound verbs derived by incorporating a noun from an adverbial phrase regularly form their passives in /−tia/. (6) In general, /−tia/ can be used when the conventional passive termination for a given verb is not remembered. (pp. 592–593)

Let us look, for instance, at (2). While the verb stem [mau] ‘to carry’ takes an [r] in the passive form [mauria], the corresponding causative form [whakamau] ‘to cause to carry’ takes the passive form [whakamautia], and not *[whakamauria]. If the different consonants are to be attributed to different endings on the verb stem, then there is no way to explain why hypothetical /maur/ should not be realized as *[whakamauria] in the passive causative.

It thus appears that there are different classes of suffixes with different initial consonants, and that the correct suffix is chosen with respect either to the verb stem or to a particular grammatical category (for example, causative passive). Since this solution will require diacritics on verb stems, it can be predicted that the difficulty of learning such a language will lead to eventual levelling out of the different consonant classes of suffixes. It is concluded, then, that the correct solution is a complex one, and not one which any present conception of a feature-counting simplicity metric would predict.