of phonological features is identical to the inventory of phonetic features, and that languages implement these universal phonetic features in various linguistic ways. In other words, phonetic features can be "phonologized" by individual languages. Of course, it may be that a phonetic feature is used phonologically by one language but not by another. In stating phonological rules, features will be chosen which seem to best explain the motivation of the processes in question.

3

PHONOLOGICAL ANALYSIS

3.0 Different Views of the Phoneme

In Chapter 1 the difference between phonetics and phonology was discussed. It was shown that in some cases phonological representations are not identical to phonetic transcriptions. In addition, the notion of distinctiveness was discussed in Chapters 1 and 2. It was claimed, for instance, that two languages can have exactly the same inventory of phonetic sounds (or phones), but significantly different phonological systems. That is, the same sounds can be organized in different ways. Just how much emphasis is to be given to these "different ways" is a matter of much debate, as we shall see.

In this chapter we shall examine the nature of phonological analysis. Since phonologists disagree in their basic assumptions about the nature of phonology, we shall see that the specific analysis of the phonetic data of a language greatly depends on the phonological *theory* underlying the analyst's work, a fact which must be constantly kept in mind. All phonologists agree that it is necessary to recognize both phonetic units (phones) and phonological units (phonemes). But there are many differences beyond this basic agreement. In 1.3 the *phoneme* was defined as a minimal unit of sound capable of distinguishing words of different meanings. Both /p/ and /b/ are phonemes in

English, because they are capable of making a meaning difference, as in the words *pin* and *bin* or *cap* and *cab*. The exact interpretation of the fact that the word *pin* means something different from the word *bin* depends crucially on one's conception of what a phoneme is.

In the following sections we shall present three views of the phoneme. In **3.1** we shall see that some linguists (particularly in America in the 1940s and 1950s) attempted to assign sounds to phonemes on the basis of their distributional properties. In **3.2** we shall see that other linguists (particularly those of the Prague School in Europe in the 1930s) assigned sounds to phonemes on the basis of their functioning within a system of oppositions. Finally, in **3.3** it will be seen that a third group of linguists view the phoneme as a psychological sound unit. Each of these approaches has provided insights into the nature of phonology, and the discussion will, hopefully, provide a historical perspective.

3.1 The Phoneme as a Phonetic Reality

The first view asserts that the phoneme represents a physical phonetic reality. That is, sounds which belong to the same phoneme share important phonetic properties. Thus Daniel Jones (1931:74) defines the phoneme as "a family of sounds in a given language, consisting of an important sound of the language together with other related sounds, which take its place in particular sound-sequences." Similarly, Gleason (1955:261) defines the phoneme as "a class of sounds which: (1) are phonetically similar and (2) show certain characteristic patterns of distribution in the language or dialect under consideration." Under this view the phoneme is seen as a convenient label for a number of phonetic units. Thus /p/ may stand for [p], [p^h], [p:], [p'], etc.

3.1.1 Minimal Pairs

The major task, then, for a phonologist holding this view of the phoneme is to determine which sounds belong in the same class. In order to do this, it is necessary to examine the *distribution* of the sounds in question. If two sounds which are phonetically similar occur in the same phonetic environment, and if the substitution of one sound for the other results in a difference in meaning, then these sounds are assigned to *different* phonemes. Thus, to continue the same example, if [p^h] is substituted for the [b] in *bin*, a different word results (namely *pin*). On the other hand, if [p^h] is substituted for the [p] in *spin* (see Chapter 1), we do not obtain a different word but rather just a slightly distorted mispronunciation, which is likely to be interpreted as [spɪn] in any case. We conclude that [p^h] and [b] belong to different phonemes, while [p^h] and [p] belong to the same phoneme.

It can easily be demonstrated that two sounds belong to different phonemes if we find two words which differ only in that one word has one of these two sounds in a given position (for example, at the beginning of the word), while the other word has the other sound in the same position. Two such words, which differ only by one sound, are said to constitute a *minimal pair*. Thus *pin* and *bin* are a minimal pair, since they differ only in their initial consonant, just as *cap* and *cab* are a minimal pair, since they differ only in their final consonant. On the other hand, *pin* and *bit* do not constitute a minimal pair, since they differ in both their initial *and* final consonants. Finally, *pin* and *nip* are not a minimal pair, since, although they involve the same three sounds, there are actually *two* differences between these two words: initially, *pin* has [p^h] while *nip* has [n], and finally, *pin* has [n] while *nip* has [p^h] (pronounced alternatively as an unreleased [p]).

We thus conclude that whenever we can establish a minimal pair, the two different sounds are phonetic manifestations of two different phonemes. The above examples involving *pin*, *bin*, and *spin* are consistent with our earlier definition of the phoneme as a *minimal unit of sound capable of making a meaning difference*. The sounds [p^h] and [p] do not make a meaning difference in English, and so we assign them to the same phoneme, let us say /p/. This phonological unit, on the other hand, contrasts with the [b] in *bin*, and this latter sound must therefore be assigned to a different phoneme, namely /b/. The following minimal pairs illustrate the pervasiveness of the opposition between /p/ and /b/ in English:

INITIAL	MEDIAL	FINAL
pin	rapid	rip
bin	rabid	rib

It should be noted, relevant to the discussion in Chapter 2, that establishing such minimal pairs reveals what the distinctive features of the language are. Thus, we can see from the above examples that voicing is distinctive in English. Such minimal pairs as *tin* : *din* and *c*[k]*ot* : *got* reveal the distinctiveness of voicing in other consonantal oppositions.

3.1.2 Complementary Distribution

The existence of minimal pairs facilitates the work of the linguist seeking to establish phonemic contrasts in this way. As Hockett (1955:212) puts it: "minimal pairs are the analyst's delight, and he seeks them whenever there is any hope of finding them." This implies that it is not always possible to find minimal pairs, and this may be due to a variety of factors. It may simply be an accident that a language does not have in its vocabulary a

minimal pair which distinguishes two sounds which theoretically could in fact be found in exactly the same position. In such cases it is necessary to rely on "near-minimal pairs." The German words *Goethe* [gø:tə] and *Götter* [gœtər] 'gods' are a near-minimal pair for the vowel phonemes /ø:/ and /œ/. They differ not only in their first vowel, but also by the presence vs. absence of a final /r/ phoneme. However, one can assume that the final /r/ of *Götter* is not likely to have an influence on the first vowel—and can therefore be disregarded in assigning [ø:] and [œ] to different phonemes.[1]

There is, however, sometimes a structural reason why two sounds cannot occur in the same environment. We have already seen, in Chapter 1, that the sounds [p^h] and [p] are generally not found in the same environment. Since this is the case, it will be impossible in English to find a minimal pair in which one word differs from the other only in that it has [p^h] instead of [p]. When two sounds are found in different environments, this is termed *complementary distribution*; the two sounds are found in mutually exclusive environments.

These environments may be stated in terms of syllable, morpheme, or word structure or in terms of adjacent segments. An example involving both comes from standard Spanish dialects. Although the words *saber* 'to know,' *nada* 'nothing,' and *lago* 'lake' are written with *b*, *d*, *g*, they are pronounced respectively [saβer], [naða], and [laɣo], that is, with the voiced nonstrident fricatives [β], [ð], and [ɣ]. On the other hand, these letters are pronounced [b], [d], and [g] in the words *banca* 'bench,' *demora* 'delay', and *gana* 'desire.' If one were to look closely at the facts of Spanish, one would discover that the sounds [β, ð, ɣ] are in complementary distribution with the sounds [b, d, g]. While the details are somewhat more complicated (see Harris, 1969:38–40), in these examples voiced stops appear at the beginning of a word, while voiced fricatives appear between vowels. That it is the intervocalic environment that is conditioning the voiced fricatives is seen from the following examples:

la banca	[la βaŋka]	'the bench'
la demora	[la ðemora]	'the delay'
la gana	[la ɣana]	'the desire'

When one adds the feminine definite article *la*, the voiced stops are then in intervocalic position (that is, between vowels), and must therefore "spirantize" to become [β, ð, ɣ]. Since these voiced fricatives (or spirants) are in complementary distribution with the voiced stops, we have only one series of phonemic consonants and not two. In a phonemic analysis based on the

[1] While the vowel of *rib* is actually longer than that of *rip* (see **5.2.5**), thereby disqualifying [rɪ:b] and [rɪp] as a true minimal pair, it is often necessary to factor out such low-level phonetic detail in phonemic analysis.

distribution of sounds, [b] and [β] would be said to be *allophones* of the same phoneme /b/, just as [d] and [ð] are allophones of /d/, and [g] and [ɣ] allophones of /g/. An allophone is, then, a phonetic realization of a phoneme in a particular environment. The voiced fricative [β] is the allophone of the phoneme /b/ found between vowels, just as the voiced stop [b] is the allophone of /b/ found at the beginning of a word.

In more recent approaches to phonology, such statements of allophonic distributions have been superseded by the explicit formulation of phonological *rules*. Thus, a rule such as the following,

$$\begin{bmatrix} b \\ d \\ g \end{bmatrix} \rightarrow \begin{bmatrix} \beta \\ \eth \\ \gamma \end{bmatrix} / V \_\_ V$$

would be postulated for Spanish, by which underlying (or phonemic) /b, d, g/ are converted to [β, ð, ɣ] between vowels. In terms of distinctive features, this rule would be formulated as follows:

$$\begin{bmatrix} +\text{voice} \\ -\text{nasal} \end{bmatrix} \rightarrow [+\text{cont}] / [+\text{syll}] \_\_ [+\text{syll}]$$

An oral voiced consonant becomes continuous (that is, a fricative) between vowels (see **4.3.1.2** for the abbreviatory conventions used in this rule).

Another case of complementary distribution comes from Standard German. Note the distribution of the fricatives [ç] and [x] in the following German words (see **1.4**):

siech	[zi:ç]	'sickly'	*Buch*	[bu:x]	'book'
mich	[mɪç]	'me'	*hoch*	[ho:x]	'high'
Pech	[pɛç]	'pitch'	*noch*	[nɔx]	'still'
		Bach [bax]	'brook'		

The velar fricative [x] appears after the back (and rounded) vowels [u:, o:, ɔ], as well as after the central (unrounded) vowel [a]. The palatal fricative [ç] is found after front (palatal) vowels, including front rounded vowels, for example, *Bücher* [bü:çər] 'books.' Since the central vowel [a] is specified [+back] in distinctive feature theory (see **2.3.3.2**), this complementary distribution is based on the distinction between preceding front and back vowels. Notice that it also extends to the diphthongs written *ai/ei*, *eu/äu*, and *au—reich* [raɪç] 'rich,' *räuchern* [rɔɪçərn] 'to smoke (meat),' *Rauch* [raʊx] 'smoke.' Since plural formation in German often involves the fronting (or *umlauting*) of a vowel, there will be numerous nouns with [x] in the singular (after a back vowel) and [ç] in the plural (after a front vowel). In addition to the alternation between [x] and [ç] seen in *Buch* and *Bücher* above, other examples are *Dach* [dax] 'roof,' pl. *Dächer* [dɛçər], and *Loch* [lɔx] 'hole,'

pl. *Löcher* [lœçər]. The palatal fricative [ç] is therefore an allophone of the phoneme /x/ after front vowels, as stated in the following rule:

$$x \rightarrow ç \;/\; [-back] \_\_ \underset{V}{}$$

Since only [ç] can occur after a consonant, for example, *Storch* [stɔrç] 'stork,' or at the beginning of a word, for example, *Chemie* [çemi:] 'chemistry,' the exact distribution of [x] and [ç] is somewhat more complicated than the above rule would indicate.

3.1.3 Phonetic Similarity

While complementary distribution is generally a clue to the phonological analysis of a language, there are cases where one might wish to maintain phonemes in complementary distribution. That is, it may be necessary to view some sounds in complementary distribution as belonging to separate phonemes. One well-known case concerns the distribution of [h] and [ŋ] in English. As seen in such words as *head, heart, enhance,* and *perhaps,* [h] occurs only at the *beginning of a syllable* (*enhance* and *perhaps* are syllabified as *en-hance* and *per-haps*). On the other hand, as seen in such words as *sing* [sɪŋ], *singer* [sɪŋ-ər], and *finger* [fɪŋ-gər], [ŋ] always occurs at the *end of a syllable.* Just as there are no English syllables ending in [h], there are no English syllables beginning with [ŋ]. It would thus appear that [h] and [ŋ] are in complementary distribution and should therefore, as suggested in **3.1.2,** be assigned as allophones of the same phoneme.

While we shall ultimately argue that [ŋ] should be recognized as the phonetic reflex of a phonemic /ng/ sequence (see **3.3.1**), let us ignore this analysis for the time being. A solution which would assign [h] and [ŋ] to the same phoneme would appear unsatisfactory to most phonemicists, since the two sounds appear to have very little in common. While [pʰ] and [p] are both voiceless labial stops in English, just as [b] and [β] are both voiced labial obstruents in Spanish, [h] and [ŋ] have little more in common than that they are both consonants. [h] is voiceless, while [ŋ] is voiced; [h] is a fricative, while [ŋ] is a (nasal) stop; [h] is oral, while [ŋ] is nasal; [h] is glottal, while [ŋ] is velar, etc. In order to rule out a solution which would assign these two sounds to the same phoneme, one must appeal to the notion of *phonetic similarity.* As Hockett (1942:103) puts it, "if *a* and *b* are members of one phoneme, they share one or more features."

The whole question of phonetic similarity is a complex one. In particular, it is not quite clear whether this criterion for assigning sounds to the same phoneme means that these sounds must share a phonetic property not shared by other sounds or simply that they must share a phonetic property. A good example comes from Gwari (Hyman, 1972a:190). The phoneme /l/ is realized as a voiced palatal stop /ɟ/ before /i/, /e/, and /y/. Thus, /lí/ 'to eat' is pronounced [ɟí] and written orthographically as *gyí.* On the other hand, the

phoneme /g/ is realized as [ɟ] before /i/ and /e/. It seems clear that the palatal stop (which is a realization of the phoneme /l/) is more phonetically similar to [g] (as the main allophone of /g/) than is [ɟ], and yet it is [ɟ] and not [ɟ] which belongs to the /g/ phoneme. Thus, while allophones share constant phonetic properties, there is no way of assigning sounds to phonemes on this basis alone. Since we shall argue for the psychological reality of phonemes in **3.3,** we can restate this problem in the following terms: while allophones of the same phoneme share phonetic properties, it is not possible to determine which sounds speakers of a language will judge as most similar by means of examining the phonetic data alone. Instead, it is necessary to evaluate the phonetic data on the basis of the entire phonological system, as will be seen in **3.2.**

3.1.4 Free Variation

Thus far we have discussed cases where two phones are assigned to one phoneme. In all of these cases the two allophones have been seen to be conditioned by context. For this reason they are sometimes referred to as *contextual variants* or *combinatory variants* (Trubetzkoy, 1939:49). However, it is possible that two phones may appear in the same context without causing a change in meaning. In this case they are usually analyzed as *free variants* or optional variants (Trubetzkoy, 1939:46). In English, final voiceless stops occur both aspirated and unaspirated, for example, [mæpʰ] or [mæp°] 'map,' [mætʰ] or [mæt°] 'mat.' In these words two phones are found in the same context, and no meaning difference results. We therefore cannot assign [pʰ] and [p°] or [tʰ] and [t°] to different phonemes. These differences would appear to have no effect on the establishing of phonemic contrasts, and the same speaker may sometimes use one phonetic realization of a phoneme and sometimes the other.

Recently this notion of free variation has come under attack by sociolinguists (for example, Labov, 1971:432–437). Labov points out that free variants often have sociological significance, and that these variants should be accounted for quantitatively. That is, rules should be provided which account for the relative frequency of "free variants." The same speaker may use one variant in one sociological situation, while he may use the other in another situation. A number of examples have been pointed out in the literature. For example, it is well known that some French speakers use an alveolar trill [r] when they are home in a small town or village, but a uvular fricative [ʁ] when they visit Paris. This particular example illustrates that some variants are due to sound changes which have not been uniformly diffused throughout a community. One group, which enjoys greater prestige throughout the community, may acquire one variant, while another group of lesser status may acquire another variant. When a speaker of the second group comes in contact with speakers of the first group, the result is "dialect

mixture." In some cases, however, the two forms coexist in the same dialect as the result of continued contact.

It is sometimes necessary to speak of free variation among phonemes. Thus, the difference between /i/ and /ɛ/ normally makes a meaning difference, for example, *beat* and *bet*. However, the word *economics* can be pronounced with either initial /i/ or /ɛ/, without a consequent meaning change. Similarly, although /u/ and /ʊ/ contrast in words such as *kook* and *cook*, the words *roof* and *root* can be pronounced with either of these vowels. It is therefore possible not only to have noncontrasting allophones in the same context but also to have noncontrasting phonemes in the same context in isolated words.

3.1.5 Discovery Procedures

A number of American linguists of the 1940s and 1950s, who held the view that the phoneme should be defined as a class of sounds, attempted to provide a methodology or set of discovery procedures for establishing phonemes. Harris (1951) devotes several chapters to the way phonemic analysis should be done, but avoids a general theoretical statement as to what the concept of the phoneme represents (for example, is it psychologically real in the sense of **3.3**). Pike (1947a:63) succinctly defines the phoneme as follows: "a *phoneme* [his emphasis] is one of the significant units of sound arrived at for a particular language by the analytical procedures developed from the basic premises previously presented." Similarly, Hockett (1942:100) defines the phoneme as "a class of phones determined by six criteria." These criteria, which are treated in **3.4**, include similarity, nonintersection (that is, no phonemic overlapping), contrastive and complementary distribution, completeness, pattern congruity, and economy. In the writings of such linguists, as argued by Chomsky (1957, 1964), emphasis is placed on the way a language should be analyzed, rather than on the way a language *is*. While most theorists have been concerned with whether the phoneme represents a phonetic reality, a phonological reality, or a psychological reality (as discussed in this chapter), it is possible to avoid the question of what the phoneme is and ask only whether a given sound belongs to one or another phoneme. Consistent with this approach is Twaddell's argument (1935) that the phoneme should be regarded as a convenient fictitious unit whose reality is yet to be proven. Chao (1934:38) on the other hand, states: "given the sounds of a language, there are usually more than one possible way of reducing them to a system of phonemes, and . . . these different systems or solutions are not simply correct or incorrect, but may be regarded only as being good or bad for various purposes." One such purpose, for instance, is clearly stated by Jones (1931:78): "The main object of grouping the sounds of a language together into phonemes is to establish a simple and adequate way of writing the language." In stating the goal of phonemic analysis as such, Jones has reduced the discussion of what a phoneme is or represents to the question of how one can best write a language phonemically. As we shall see in **3.2** and **3.3**, other linguists have asked more of their phonemes.

3.2 The Phoneme as a Phonological Reality

The definition of the phoneme in purely phonological terms is characteristic of the Prague School. Trubetzkoy (1939:36) defines the phoneme as "the sum of the phonologically relevant properties of a sound." For him, phonemes are defined in terms of *oppositions* in a phonological system. The important notion in Prague School phonology is "function": "The phoneme can be defined satisfactorily neither on the basis of its psychological nature [see **3.3**] nor on the basis of its relation to the phonetic variants, but purely and solely on the basis of its function in the system of language" (Trubetzkoy, 1939:41). Thus, a phoneme is a minimal unit that can *function* to distinguish meanings. It is not a sound or even a group of sounds, but rather an *abstraction*, a theoretical construct on the phonological level. It is defined in terms of its contrasts within a system. For example, we saw in Chapter 1 that the /b/ phoneme in English is very different from the /b/ phoneme in Berber, since in the latter case there is no /p/ to contrast with. Approaching the phoneme as a class of sounds, one would miss the fact that although [b] is assigned to /b/ in both languages, there is a basic difference between this phoneme in English and in Berber.

3.2.1 Phonemic Overlapping

In several of the examples discussed, two phones were assigned to the same phoneme, for example, [x] and [ç] in German. One issue which reveals a fundamental difference between defining the phoneme as a class of sounds and defining it by its function within a phonological system of oppositions is the question of whether one phone can be assigned sometimes to one phoneme and at other times to another phoneme. Such a possibility, termed *phonemic overlapping*, is raised by Bloch (1941) and is discussed by a number of European phonologists (for example, Martinet, 1947; Fischer-Jørgensen, 1956:591). An example discussed by Jakobson, Fant and Halle (1952:5) concerns Danish /t/ and /d/. In syllable-initial position these phonemes are pronounced, respectively, [t] and [d], for example, [tag] 'roof' and [dag] 'day.' In syllable-final position, however, /t/ is pronounced [d] and /d/ is pronounced [ð], as seen in the following words:

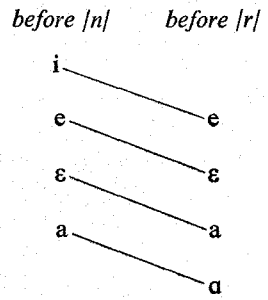
/hat/ → [had] 'hat'
/had/ → [hað] 'hate'

We must recognize for Danish a rule which "weakens" consonants in syllable-final position. The result is that the [d] of 'day' must be assigned to the phoneme /d/, but the [d] of 'hat' must be assigned to the phoneme /t/. Thus, one phone is assigned to one of two phonemes, depending on the context.

Such examples of overlapping pose a problem for adherents of the phonetic similarity criterion in phonemic analysis. What it means is that it is not possible to predict what phoneme a given phone will be assigned to on the basis of its phonetic character alone, since we have seen [d] to be assigned once to /t/ and once to /d/. The idea that phones and phonemes could be identified on a one-to-one basis, that is, that a given sound will always belong to a given phoneme and a given phoneme will always be associated with a given sound, is termed *biuniqueness* by Chomsky: "the biuniqueness condition ... asserts that each sequence of phones is represented by a unique sequence of phonemes, and that each sequence of phonemes represents a unique sequence of phones" (1964:94). If one were to adhere to phonetic similarity as an overriding principle in assigning phones to phonemes, one would be forced to say that syllable-final [d] is a realization of the phoneme /d/, and that syllable-final [ð] is the realization of a third phoneme /ð/, which is found only in this position.

Just as Chomsky showed that it is necessary in phonological analysis to allow for phonemic overlapping of the kind just illustrated (and therefore argued against the biuniqueness condition), most European phonologists noted the consistency of overlapping with their view of the phonemes in a system of oppositions. Thus, Jakobson, Fant and Halle (1952:5) state: "Two patterns are identical if their relational structure can be put into a one-to-one correspondence, so that to each term of the one there corresponds a term of the other." In other words, [t] is to [d] in syllable-initial position as [d] is to [ð] in syllable-final position. In the terms of Martinet (1960:60), physical identity does not necessarily imply linguistic identity.

Examples of phonemic overlapping are not particularly difficult to find. One, from Danish again, is discussed by Martinet (1947:43). As seen in the following diagram,



there are four contrastive vowel heights in Danish. The four front unrounded vowels are normally realized (indicated in the diagram as before /n/) as [i, e, ɛ, a]. However, before /r/ a rule of vowel lowering applies, yielding the phonetic series [e, ɛ, a, ɑ]. While this process has modified the phonetic characteristics of each vowel phoneme, it can easily be seen that the relation between the four vowels has remained constant. Thus, the vowel [e] of [er] sequences is assigned to the /i/ phoneme, even though the vowel [e] of [en] sequences is assigned to the /e/ phoneme. Danish is analyzed in this way because the phoneme /i/ is defined not in *phonetic* terms but rather in terms of its *function* within the total vowel system. In particular, rather than defining /i/ as consisting of a particular class of sounds, we define /i/ as the highest front vowel in Danish. Similarly, we define /e/ as the second-highest front vowel. Thus, when we have to assign the [e] of [er] sequences to a vowel phoneme, we choose /i/, since [e] here represents the highest front vowel before /r/. As in the case of Danish /t/ and /d/, we can apply Jakobson, Fant and Halle's notion of relational structure: [i] is to [e] before /n/ as [e] is to [ɛ] before /r/.

3.2.2 Neutralization

Bloch (1941:66–67) makes the distinction between *partial overlapping* and *complete overlapping*: "The intersection or overlapping of phonemes will be called partial if a given sound x occurring under one set of phonetic conditions is assigned to phoneme A, while the same x under a different set of conditions is assigned to phoneme B; it will be called complete if successive occurrences of x under the same conditions are assigned sometimes to A, sometimes to B." The two examples discussed in the preceding section both represent cases of partial overlapping.

A case of complete overlapping pointed out by Bloch involves English /t/ and /d/. Intervocalically, /t/ and /d/ are both pronounced as an alveolar tap [ɾ]. Thus, for many speakers of American English, the words *betting* and *bedding* are pronounced identically, that is, as [bɛɾɪŋ]. One might, however, attempt to assign different phonemic representations to the two words on the basis of the fact that *betting* contains the word *bet* and *bedding* contains the word *bed*. Assuming that the velar nasal should be phonemicized as /ng/ (see **3.3.1**), the two phonemic representations would then be /bɛtɪng/ and /bɛdɪng/. In this case, however, it would be necessary to state that both /t/ and /d/ have the allophone [ɾ] in the *same* environment, namely in intervocalic position. What this means in terms of Prague School phonology (see **2.2.3**) is that an opposition has been *neutralized* in this position.

While /t/ and /d/ contrast initially, as in the words *tin* and *din*, and while they contrast finally, as in the word *bet* and *bed*, they do not contrast intervocalically (with the additional restriction that the preceding vowel is stressed).

In **2.2.3**, such an opposition was termed *neutralizable*. On the other hand, the contrast between /p/ and /b/ is, at least with respect to initial, medial, and final position, a *constant* opposition (see, however, footnote 3, Chapter 2). Trubetzkoy (1939:78) differentiates *positions of neutralization*, where the neutralization takes place, and *positions of relevance*, where the opposition is realized phonetically. Thus, in the above example, the intervocalic position is the position of neutralization, while the initial and final positions are the positions of relevance.

Notice that if phonemic forms such as /bɛtɪng/ are to be permitted for English, then the phonological analysis will be possible only if the phonologist goes beyond the phonetic data. In particular, it must be known whether there is a word *bet* that exists independently, and whether this word exists as a morpheme in a word such as *betting*. This consideration clearly goes beyond the distributional analysis inherent in discovering complementary distribution. In this case we not only need to know whether two forms are the same (one phoneme) or different (two phonemes), but also we must establish exactly which morpheme (for example, *bet* or *bed*)) is present. In otherwords, we must introduce *grammatical* information into the phonological analysis. In terms of the positions outlined in **3.3.2**, this amounts to "mixing levels."[2]

To combat the problem of neutralization, Prague School phonologists introduced the *archiphoneme*. Consider a language such as Fe?fe?–Bamileke, which has the following sequences:

	ku	či		ču
ke	ko	če		čo
ka			ča	

Since both [k] and [č] are found before /e/, /a/, /o/, and /u/, we conclude that they belong to separate phonemes, that is, /k/ and /č/. However, there is a problem concerning the vowel /i/, since only [č] is found before it. There are sequences of [či] in this language, but *[ki] is not found. If we were to analyze [či] as /či/ phonemically, Praguians would point out that this /č/ is not the same as the /č/ found in other positions. The phoneme /č/ is defined in part by the fact that it stands in opposition to /k/. Before /i/, however, this part of the definition is destroyed, since the difference between [k] and [č] cannot be used here to make a meaning difference.

Instead of calling [č] before /i/ another instance of /č/, a separate phonological unit is set up which is neither /č/ nor /k/, but which consists of all of the phonological properties *shared* by /č/ and /k/. This unit, termed an archiphoneme, is by convention written as a capital letter, here /K/. /K/ stands for a voiceless noncontinuant, which would be specified in terms of

[2] In a phonemic analysis emphasizing the distributional properties of sounds, it would probably be necessary to recognize a third phoneme /ɾ/, because of its unique distribution (it occurs only intervocalically and after certain sonorant consonants, e.g., *party* [pɑɾɾi]).

distinctive features as [+high] (that is, either palatal or velar), but which would not be specified with respect to backness. In other words, its specifications would be as seen below, with [0 back] indicating that this feature is irrelevant (left blank), since it is neutralized:

$$/K/ \; : \; \begin{bmatrix} +\text{cons} \\ -\text{syll} \\ -\text{son} \\ +\text{high} \\ 0\ \text{back} \\ -\text{ant} \\ -\text{cor} \\ -\text{voice} \\ -\text{cont} \\ -\text{nas} \\ 0\ \text{strid} \\ 0\ \text{del rel} \end{bmatrix}$$

In addition to [0 back], the features Strident and Delayed Release are not specified, since the archiphoneme does not specify whether the noncontinuant is a stop or an affricate.

Since [č] before /i/ represents the neutralization of the opposition between /k/ and /č/, it would be incorrect in this framework to phonemicize [či] as /či/. Trubetzkoy (1939:78) draws support for this approach from linguistic performance: "In neutralizable distinctive oppositions perception fluctuates: in positions of relevance both opposition members are clearly distinguished; in positions of neutralization, on the other hand, it is often not possible to indicate which of the two had just been produced or perceived." Phonemes which participate in neutralizations are thus felt by speakers to be closely related. We might presume, as a result, that speakers of the above language will regard /č/ and /k/ as more closely related than they will /č/ and /t/.

An example of neutralization often cited in the literature was discussed in **2.2.3**. In Standard German, voiced obstruents are devoiced syllable-finally. While the phonemes /t/ and /d/ contrast initially (for example, *Tier* [ti:r] 'animal' vs. *dir* [di:r] 'to you') and intervocalically (for example, *leiten* [laɪtən] ' to lead' and *leiden* [laɪdən] 'to suffer'), there is no possible contrast syllable-finally. Thus, the words *Rat* 'advice' and *Rad* 'wheel,' which are written differently, are both pronounced [ra:t]. Notice, however, that in the plurals, where a suffix is added (which causes a vowel change as well), the contrast been /t/ and /d/ resurfaces: *Räte* [rɛ:tə] 'advices' and *Räder* [rɛ:dər] 'wheels.' The question is how the final [t] of *Rat* and *Rad* should be analyzed.

Phonologists maintaining a definition of the phoneme as a class of phonetically similar sounds often disallowed complete overlapping (neutralization) and were therefore forced to analyze both 'advice' and 'wheel' as

/ra:t/. Prague School phonologists, who saw the phoneme as a unit in a system of oppositions, could not analyze the final stop of these words as /t/, since, unlike its counterpart in initial and intervocalic position, it cannot stand opposed to /d/. Therefore, an archiphoneme would be set up. As stated by Trubetzkoy, "In German the bilateral opposition d–t is neutralized in final position. The opposition member which occurs in the position of neutralization from a phonological point of view is neither a voiced stop nor a voiceless stop but the 'non-nasal dental occlusive in general'." Thus the underlying representation of both 'advice' and 'wheel' is /ra:T/, where /T/ is specified [0 voice], representing an archiphoneme sharing all of the properties common to /t/ and /d/. The words *Rat* and *Rad* thus end in a dental stop which is *redundantly* voiceless.

3.3 The Phoneme as a Psychological Reality

The original mentalist position, as espoused by Badouin de Courtenay, defined the phoneme as "a mental reality, as the intention of the speaker or the impression of the hearer, or both" (Twaddell, 1935:56). Since each time a speaker pronounces the sound [p] it is acoustically never quite the same as the last [p], the speaker must have internalized an image or idealized picture of the sound, a target which he tries to approximate. Badouin de Courtenay spoke of the phoneme as "a sound imagined or intended, opposed to the emitted sound as a 'psychophonetic' phenomenon to the 'physiophonetic' fact" (Jakobson and Halle, 1956:11). Thus, according to the argument, in Nupe (where /si/ is realized as [ši]), when a speaker pronounces [ši] 'to buy,' his real intention or abstract image is /si/. Similarly, when a speaker of American English says [aɪ mɪšə] 'I miss you,' his real intention is [aɪ mɪs yu], and so forth.

This view of the phoneme as a psychological unit was subject to attack by phonologists holding the views of the phoneme discussed in **3.1** and **3.2**. The following statement of Twaddell (1935:57) is perhaps representative of American reaction against mentalistic definitions of the phoneme: "Such a definition is invalid because (1) we have no right to guess about the linguistic workings of an inaccessible 'mind,' and (2) we can secure no advantage from such guesses. The linguistic processes of the 'mind' as such are quite simply unobservable; and introspection is notoriously a fire in a wooden stove."

Representative of the Praguian reaction to Courtenay, Trubetzkoy (1939:38) states: "Reference to psychology must be avoided in defining the phoneme, since the latter is a linguistic and not a psychological concept."

To Trubetzkoy, the phoneme is a characteristic of the linguistic system, and not of the minds of speakers:

> The fact that the concept "phoneme" is here [in Courtenay's writings] linked with such vague and nondescript notions as "psyche," "linguistic consciousness," or "sensory perception" cannot be of help in clarifying the phoneme concept. If this definition were to be accepted, one would never know in an actual case what to consider a phoneme. For it is impossible to penetrate the "psyche of all members of a speech community" (especially where extinct languages are involved). (1939:39)

Although perhaps most phonologists reacted to the strong psychological wording of Courtenay's pioneering work, this does not mean that they completely refrained from discussion of psychological (for example, perceptual) aspects of the phoneme. Virtually all theorists agree that the phonemic system of a language exerts a behavioral effect on its speakers. Few phonologists fail to make some remark about the role of the phonemic system in the perception of foreign sounds. In the words of Trubetzkoy,

> The phonological system of a language is like a sieve through which everything that is said passes.... Each person acquires the system of his mother tongue. But when he hears another language spoken he intuitively uses the familiar "phonological sieve" of his mother tongue to analyze what has been said. However, since this sieve is not suited for the foreign language, numerous mistakes and misinterpretations are the result. The sounds of the foreign language receive an incorrect phonological interpretation since they are strained through the "phonological sieve" of one's own mother tongue. (1939:51–52)

Even Harris (1954:36), who devoted so much attention to distributional analysis, wrote: "Clearly, certain behaviors of the speakers indicate perception along the lines of the distributional structure, for example, the fact that while people imitate nonlinguistic or foreign-language sounds, they *repeat* [his emphasis] utterances of their own language." While the antimentalist phonologists of the 1930–1950 era were quick to reject all psychological terminology, they did not refrain from pointing out that their nonpsychological phonemic systems have psychological validity for speakers.

The classic article on the psychological reality of phonemes is Sapir's (1933) article bearing exactly this title. In this article Sapir reports the following anecdote:

> When working on the Southern Paiute language of southwestern Utah and northwestern Arizona I spent a little time trying to teach my native interpreter ... how to write his language phonetically.... I selected *pa:βah*.... I instructed Tony to divide the word into its syllables and to discover by careful hearing what sounds entered into the composition of each of the syllables.... To my astonishment Tony then syllabified *pa:*, pause, *pah*. I say "astonishment" because I at once recognized the paradox that Tony was not "hearing" in terms of the

actual sounds (the voiced bilabial *β* was objectively very different from the initial stop) but in terms of an etymological reconstruction: *pa:* 'water' plus postposition **—pah* 'at.' The slight pause which intervened after the stem was enough to divert Tony from the phonetically proper form of the postposition to a theoretically real but actually nonexistent form. (pp. 23–24)

What this means is that Tony had knowledge of the underlying /p/ in the postposition 'at,' which by rule becomes the voice spirant [β] intervocalically. In other words, the /p/ in the phonemic representation is psychologically real.

3.3.1 Levels of Adequacy

Examples such as the above reveal that phonology goes well beyond the systematizing of phones into phonemes. There has been much recent discussion of the goals of phonology. Chomsky (1964:62ff), for example, distinguishes phonological analyses which are *observationally adequate* from those which are *descriptively adequate*. A phonological analysis is observationally adequate if it accurately transcribes the data and nothing more. It is descriptively adequate if, in addition to transcribing the data, it accounts for the knowledge (which Chomsky refers to as *linguistic competence*) of the native speaker. Let us say, for instance, that one description of English phonology states that there is a word *play* [ple] and a word *clay* [kle], but no word **tlay* (presumably to be pronounced [tle]). Such a description reaches the level of observational adequacy, since it correctly states that certain forms are observed while other forms are not.

This description cannot be said to reach the level of descriptive adequacy, however, unless it accounts for the fact that *tlay* not only is not observed but could not be a possible word in the language. The native speaker intuitively knows that it is not possible to have a [tl] cluster at the beginning of a word in English. Thus, a related fact is that English has the words *pluck* [plək] and *cluck* [klək], but no word **tluck* [tlək]. A descriptively adequate phonological description of English must include numerous constraints on consonant sequences (see **1.4.1**). Much more will be said about such constraints. For the moment it is important only to note that native speakers have knowledge of these constraints. Greenberg and Jenkins (1964) have experimentally demonstrated the native speaker's ability to judge nonexistent forms for their well-formedness, both in terms of sequences which do or do not "sound English" and in terms of the distance of such forms from good English-sounding words.

An example of a phonological analysis reaching the level of descriptive adequacy concerns the velar nasal consonant [ŋ] in English. Many phonologists have observed that the velar nasal, which is written *ng* as in *sing* [sɪŋ], does not occur word-initially in English, although [m] and [n] do. A phonological analysis of English could merely state this constraint, but there

is good reason to believe that such an analysis remains too superficial. In particular, once this constraint is stated, one might further ask *why* there is such a constraint to begin with. We might hypothesize that the sound is too difficult to pronounce in this position, but then there are many languages which do in fact allow [ŋ] word-initially, as the spelling of the Vietnamese name *Nguyen* suggests. Thus, while [ŋ] is difficult for an English speaker to pronounce at the beginning of a word, its absence in this position in English cannot be explained in universal terms.

Rather, the reason we do not find word-initial [ŋ] is that it derives historically from an earlier *[ŋg]. Thus, the reason we find words such as *meat* [mit] and *neat* [nit], but not **ngeat* [ŋit], is a *historical* one. The velar nasal derives historically from [ŋg] at a stage where there was not only no word *ngeat [ŋgit] but also no word **mbeat* [mbit] or **ndeat* [ndit]. That is, a word could not begin with a nasal consonant followed by a voiced stop. What is interesting is that although the [g] of *[ŋg] has dropped, [ŋ] continues to function *as if* there were a [g] after it.

In fact, Sapir (1925:19) proposed that the sound [ŋ] be analyzed phonologically in English as /ŋg/:

> In spite of what phoneticians tell us about this sound (*b:m* as *d:n* as *g:ŋ*), no naïve English-speaking person can be made to feel in his bones that it belongs to a single series with *m* and *n*. Psychologically it cannot be grouped with them because, unlike them, it is not a freely movable consonant (there are no words beginning with *ŋ*). It still *feels* like *ŋg*, however little it sounds like it. The relation *ant : and = sink : sing* is psychologically as well as historically correct.

Chomsky and Halle (1968:85n) propose that [ŋ] should be described phonologically as /ng/. Two rules are necessary:

1 $n \rightarrow ŋ \; / \; \_\_ \; \{k, g\}$
2 $g \rightarrow \emptyset \; / \; ŋ \; \_\_ \; \#$

Rule **1** assimilates /n/ to [ŋ] before a velar consonant, for example, /sɪnk/ becomes [sɪŋk]; Rule **2** deletes [g] after [ŋ] and before a word boundary (see **3.3.2** and **6.1.2.2** for discussion of boundaries). Thus, the full derivation of *sing* is as follows:

/sɪng/ → sɪŋg → [sɪŋ] (by rules 1 and 2)

Given this /ng/ solution, a general sequential constraint can be formulated: in English, no words begin with *mb*, *nd*, or *ng*, that is, no word begins with a nasal consonant followed by another consonant. It is this constraint on the *phonological* level which explains the failure of [ŋ] to occur at the beginning of English words.

The /ng/ solution reaches the level of descriptive adequacy because it provides a principled reason for the exclusion of word-initial [ŋ]. In addition,

Fromkin (1971:34–35) presents evidence from speech errors for this analysis. She reports that someone, instead of saying *Chuck Young* (the Chancellor of UCLA), said *chunk yug*. Phonetically, this represents a change from the intended [čək yəŋ] to the speech error [čəŋk yəg]. If it is assumed that [ŋ] is phonologically /ng/, then this error (as well as others) can be explained by saying that the nasal consonant was transferred to the first word, thereby leaving a [g] sound stranded in the second word. The possibility of using data from speech errors to help choose among competing analyses seems very promising.

3.3.2 Grammatical Prerequisites to Phonology

One of the basic disagreements in the history of phonology has to do with what is referred to as "mixing levels." According to some phonologists, a phonological analysis would have to be justified on the basis of the phonetic variants alone. In particular, information from a grammatical level (that is, morphology, syntax) could not be used to justify an analysis. Hockett (1942:20–21) sums up this position: "There must be no circularity; phonological analysis is assumed for grammatical analysis, and so must not assume any part of the latter. The line of demarcation between the two must be sharp." This position was sometimes maintained by phonologists focusing on discovery procedures (see **3.1.5**). Procedures were developed by which sounds could be assigned to phonological units (phonemes), which in turn could, by other procedures, be assigned to grammatical units (morphemes, words).

We have already mentioned Chomsky's criticism (1957:50–53) of discovery procedures. However, all one needs to disprove the claim that phonological analysis can be done without recourse to grammatical information is to find a language where the phonology cannot be described without reference to the grammar, where "grammar" is used as a cover term for both morphology (word structure) and syntax (sentence structure).

Such examples are not hard to find. Specifically, many phonological descriptions require information such as (1) morphological boundaries and (2) class categories, such as nouns and verbs. A good example of the latter occurs in English. It is generally assumed that part of a complete phonology of English will deal with stress phenomena. However, the placement of stress in a word is partly dependent on whether that word is a noun or verb, as seen in the following examples:

NOUN	VERB
pérvert	pervért
súbject	subjéct
cónduct	condúct

While there are exceptions (for example, *to rámble, a lament, a babóon*), some of which can be explained in terms of syllable structure and vowel tenseness (see Chomsky and Halle, 1968), the above noun and verb forms suggest a generalization: stress falls on the first syllable in nouns, but on the second syllable in verbs. Thus, for a particular set of noun–verb pairs, stress can only be accounted for with reference to grammatical information.

Another example is found in Nupe (Hyman, 1970a). In Nupe, the phoneme /s/ is pronounced [š] before /i/; for example, /sī/ 'to buy' is pronounced [šī], but /sá/ 'to cut' is pronounced [sá]. Thus, it would appear that the difference between [s] and [š] is completely redundant, since we can predict which one is found on the basis of the following vowel. Phonemic /s/ is palatalized to [š] before the front vowel /i/ (as well as before /e/ and /ɛ/). There is, however, one exception. There is a process of reduplication in Nupe which creates nouns from verbs, for example, [šī] 'to buy' becomes [šīšī] 'buying.' The vowel in the reduplicated prefix is frequently [i] (but see Hyman, 1970a:67–69 for a fuller statement; also **3.3.5**). The exception to the palatalization of /s/ to [š] before /i/ arises when a verb such as /sá/ 'to cut' is reduplicated as [sīsá] and not *[šīsá]. If we were to base ourselves entirely on the phonetics, we would be forced to say that the difference between [s] and [š] is a distinctive one, since the utterance [šī sá] (from /sī/ + /èsá/ 'to buy a chair') is also found. Thus, [sīsá] and [šī sá] would constitute a minimal pair. Such a minimal pair, which is possible only when one of the forms is a noun derived through reduplication, should not be allowed to destroy the complementary distribution of [s] and [š] in the language, which is otherwise completely general. With a minimum of grammatical information, we can still predict when we will find [s] and when we will find [š]. Nupe speakers palatalize /s/ to [š] before /i/, *except* in such cases of reduplication (see Wilbur, 1973, for theoretical discussion).

In addition to grammatical categories such as noun and verb, it is frequently necessary to refer to grammatical boundaries in phonological analyses. The boundaries which are used in phonology (see **6.1.2.2**) include the full-word boundary (##), the internal word or stem boundary (#), and the general morpheme boundary (+). An example of the relevance of such boundaries comes from Fe?fe?–Bamileke. Consider the following data:

(a)	pǒ	'hand'	mbǒ	'hands'
	pē:	'accept'	mbē:	'and accept'
	púɑ	'two'	ntām púɑ	'two hearts'
	pì:	'profit'	tūm pì:	'send the profit'
(b)	vūp	'whip'	vūbī	'whip him/her'
	ŋgǔp	'hen'	ŋgūbà	'my hen'
	pū:	'children'	pē: pū:	'accept the children'

In several of these examples there is an alternation between [p] and [b].

Let us assign [p] and [b] as allophones of the phoneme /p/ (see Hyman 1972b, Chapter 3, for discussion of this solution). In (a), /p/ is realized as [b] only in the first two examples in the right-hand column, as the result of a rule which voices /p/ after [m]:

$$p \rightarrow b / m \_\_$$

However, in the third and fourth examples in the right-hand column, voicing does not take place. The above rule is in effect *blocked* by the full-word boundary in the phrases 'two hearts' and 'send the profit.' Since there is only an internal word boundary in 'hand' and 'and accept,' that is,

/m#pə̃/	'hands'	/ntām##púɑ/	'two hearts'
/m#pē:/	'and accept'	/tūm##pì:/	'send the profit'

the rule is not prevented from applying.

Similarly, the first two examples in the right-hand column of (b) show /p/ becoming [b] intervocalically, as in the following rule:

$$p \rightarrow b / V \_\_ V$$

Since there is a full word boundary in /pē:##pū:/ 'accept the children,' no voicing takes place. On the other hand, the internal word boundary of /vāp#ī/ 'whip him/her' and /ŋgāp#à/ 'my hen' does not block the above rule. Thus the distribution of [p] and [b] in Fe?fe?–Bamileke can only be accounted for if it is possible to refer to word boundaries. Otherwise we would be forced to conclude that the difference between [mp] and [mb] is a distinctive one, necessitating the positing of two phonemes /p/ and /b/.

Although grammatical boundaries play a role in phonology, some linguists attempted to introduce "phonological" junctures in order to avoid mixing levels. The junctures are responsible for phonetic differences in such phrases as *why try* [wa:ɪ tʰraɪ] and *white rye* [waɪt raɪ]. Thus, Z. Harris points out (1951:88): "Many of the junctures set up... without reference to morphologic boundaries turn out nevertheless to come precisely at morphologic boundaries." While many of the phonologists eschewing the use of grammatical information did not follow their own advice in practice, not all of the linguists of the descriptivist era of the 1940s and 1950s in the United States were even theoretically in agreement, as is evident from the following statement made by Pike (1947b:158): "If language actually works as a unit, with grammatical configurations affecting phonetic configurations, why should we not describe the language and analyze it in that way? If forced to do so, why pretend we are avoiding it?" The consequences, however, show that one cannot proceed by operational steps from the physical sounds to the phonemes and from the phonemes to the morphemes, etc. Since no alternative hypotheses or criteria were advanced, this particular theory breaks down.

3.3.3 Morphophonemics

It is thus possible that the phonetic reflexes or realizations of phonemes not only reveal phonetically determined oppositions but also are determined by grammatical facts. We have discussed two possible solutions to the German case of final devoicing (**3.2.2**). The first solution, that characteristic of American phonemics, is to identify the phonetic shape of the segment found in the position of neutralization with the phonological representation. Thus, *Rat* and *Rad* will both be represented as /ra:t/. The second solution, that characteristic of the Prague School, is to posit an archiphoneme in the position of neutralization. Thus, German *Rat* and *Rad* are both represented as /ra:T/. Both of these solutions fail to give an explicit account of the fact that one instance of [ra:t] (let us say $[ra:t]_1$) alternates with a plural form with [t], that is, [rɛ:tə] 'advices,' while the other instance of [ra:t] (let us say $[ra:t]_2$) alternates with a plural form with [d], that is, [rɛ:dər] 'wheels.' The fact that there are basically two kinds of final *t*s in German is overlooked.

Clearly, there is a certain relationship between [t] and [d] in German. Since this relationship is missed by phonemic analysis, a separate, more abstract level is recognized, called the *morphophonemic* level, whose basic unit is the *morphophoneme*. The basic motivating principle is that it should be possible to give one representation to each morpheme (minimal meaningful unit of grammar) and derive all of the allomorphs from this one "base form" (barring, of course, the possibility that two allomorphs may not be phonologically related to one another, for example, *go* and *went*). The morpheme 'wheel' has two alternate phonemic forms or *allomorphs* in German: it has the allomorph /ra:t/ when the final consonant is followed by pause, but the allomorph /rɛ:d/ when the final consonant is followed by a vowel. This is no accident. The same could be said about the noun *Bund* [bʊnt] 'union' and its plural form *Bunde* [bʊndə]. This morpheme has the allomorph /bʊnt/ when the alveolar consonant is before pause, but the allomorph /bʊnd/ when there is a following vowel.

The base forms of these morphemes are {raT} and {bunT}, respectively. These capital letters are employed to represent morphophonemes and should not be confused with the archiphonemes discussed in **3.2.2**. Here {T} is the morphophoneme which is sometimes represented by the phoneme /t/ and sometimes by the phoneme /d/. As Z. Harris states: "Each morphophonemic symbol thus represents a class of phonemes and is defined by a list of member phonemes each of which occurs in a particular environment" (1951:225). The example he discusses concerns the alternation between /f/ and /v/ in

English, as exemplified in the forms *knife/knives*, *wife/wives*, *leaf/leaves*, *thief/thieves*, etc. For such allomorphs Harris proposes the morphophoneme {F}, for example, {naɪF} 'knife,' which is sometimes realized as the allomorph /naɪf/ (in the singular) and sometimes as the allomorph /naɪv/ (in the plural). Notice that while a word such as *thief* will have the base form {θiF} (since its plural *thieves* is formed with /v/), a word such as *chief* will have the base form {čif} (identical with its phonemic representation /čif/), since its plural is *chiefs* and not **chieves*.

3.3.4 Systematic Phonemics

This notion of one base form per morpheme is carried over into the models of generative phonology presented as early as Halle (1959) and still characterizing most of the work being done in this theory.[3]

The view is expressed in generative phonology that native speakers of a language *tacitly* know (that is, the knowledge is not necessarily conscious) that certain forms are related and that this relatedness must be captured somehow in the grammar. These phonologists propose that highly abstract systematic phonemic representations (equivalent in many respects to morphophonemic representations) be postulated, from which rules derive the various surface realizations. By postulating one underlying form at the systematic phonemic level, from which surface alternants are derived, the tacit knowledge speakers have of general or systematic relationships (termed *linguistically significant generalizations*) in the phonological structure is accounted for. Chomsky and Halle (1968) point out that, as a result of the Great English Vowel Shift, there are vowel alternations such as those seen in the following words (we shall limit this discussion to front vowels only):

[iy]	:	serene	[ɛ]	:	serenity
		obscene			obscenity
[ey]	:	profane	[æ]	:	profanity
		inane			inanity
[ay]	:	divine	[ɪ]	:	divinity
		sublime			sublimity

On the basis of these alternations (and various other arguments), Chomsky and Halle propose the following abstract systematic phonemic representations of these morphemes:

/serēn/	/profǣn/	/divīn/
/obsēn/	/inǣn/	/sublīm/

[3] For a thorough statement of the "standard model" of generative phonology, i.e., of systematic phonemics, see Chomsky and Halle (1968); for a more simplified and concise introduction, see Schane (1973a).

That is, tense vowels (indicated by $\overline{V}$) are set up. Notice how closely these underlying forms resemble English orthography. This comes as no surprise, since these abstract forms coincide with historical reconstructions, which are preserved in the orthography.

Three rules are required to produce the correct phonetic forms. First, there is a vowel laxing rule, which for our purposes applies before the *–ity* suffix.[4] Thus, /serēn/ becomes *seren* before the *–ity* suffix. Second, there is a vowel shift rule which changes /ī/ to ǣ, /ē/ to ī, and /ǣ/ to ē. Finally, there is a diphthongization rule by which ǣ becomes [æy], ī becomes [iy], and ē becomes [ey]. The derivations for [səriyn] and [sərɛnɪti] are given below:

/serēn/	/serēn+iti/	
	seren+iti	Laxing before *–ity*
serīn		Vowel shift
seriyn		Diphthongization

In Chomsky and Halle's framework, lax *i* and *e* are to be identified with phonetic [ɪ] and [ɛ], respectively. The schwa found in the words *serene* and *serenity* is due to a rule that reduces unstressed vowels to schwa.

The vowel shift rule is also used in conjunction with other alternations in the language. Chomsky and Halle point out (p. 234) that alternations such as *resign* : *resignation* and *paradigm* : *paradigmatic*, where the simple form has [ay] and the complex form [ɪg], must be accounted for, since these forms are related. This relatedness is accounted for by providing a unique base form for each morpheme. Looking at the word *resign*, Chomsky and Halle argue for the systematic phonemic representation /rē=sign/. A number of observations are relevant here. The equal sign (=) represents a special morpheme boundary which is necessary in the following rule (p. 95):

$$s \rightarrow z \,/\, V = \_\_\, V$$

The reason Chomsky and Halle wish to posit an /s/ in the underlying form is that the same morpheme, they claim, occurs in words such as *consign*, where the same = boundary is recognized. They argue that this boundary must function in the rule voicing /s/ to [z], since when there is no boundary, or when there is a full + morpheme boundary (or perhaps a word boundary #), /s/ remains [s] (for example, *reciprocate*, *re-sign* /rē#sign/ 'to sign anew').[5]

[4] This rule actually laxes the vowel of the third syllable from the end of the word. Thus, the vowel of the syllable directly preceding *–ity* will automatically become lax.

[5] There are, however, important exceptions. While *design* is pronounced with [z], as predicted by the above rule, *desist* is pronounced with [s] by some speakers. Since this word is represented underlyingly as /dē-sist/, it should undergo the same intervocalic voicing of /s/ as /rē-sist/, which is pronounced [rizɪst]; cf. *consist*, which is pronounced with [s], since /s/ is not in intervocalic position.

Addressing ourselves now to the problem of the /g/ in *resignation* and its absence in *resign*, Chomsky and Halle propose a *g*–deletion rule, the effect of which is to tense the preceding vowel. (They discuss certain possibilities, in particular an intermediate [ɣ] which tenses the preceding vowel and later drops.) Let us state the *g*–deletion as follows: /g/ falls when it occurs before a syllable–final /n/.[6] Thus, since the word–final /n/ of *resign* is also syllable–final, the /g/ falls. However, since *resignation* is syllabified as *re–sig–na–tion*, the /g/ remains. The derivation for [riyzayn] is as follows:

/rē=sign/	Underlying (systematic phonemic) form
rē=zign	Voicing of /s/
rē=zīn	Drop of /g/ with concomitant tensing
ri=zǣn	Vowel shift
[riyzæyn]	Diphthongization

(The resulting diphthong [æy] is slightly modified to [ay] (= [aɪ]) by another rule.)

3.3.5 Phonological Abstractness

It should be clear from the previous section that considerable "abstractness" is achieved by Chomsky and Halle and others in setting up underlying forms. The resulting systematic phonemic representations are considerably more distant from the surface phonetics than any other school of phonology ever would have tolerated.

Systematic phonemics, however, goes beyond proposing an abstract morphophonemic level, since, in developing this theory of phonology, Halle (1959) proclaimed the nonexistence of both the traditional phoneme and the phonemic level. That is, between the systematic phonemic level (resembling the old morphophonemic level) and the (systematic) phonetic level there would now be no linguistically significant level corresponding to the old phonemic level.

Chomsky (1964) and Postal (1968) devote much time to supporting this view. While phonology has experienced since *The Sound Pattern of English* a shift back in the direction of a less abstract phonological level (see Kiparsky, 1968a; Schane, 1971; Stampe, 1972a), it would be worthwhile to briefly examine the kind of argument given against what has come to be known as the "autonomous" or "taxonomic" phoneme (autonomous because some phonemicists refused to admit grammatical information into their phonological analysis, and taxonomic because sounds were merely classified, ignoring important phonological generalizations expressible by rule).

[6] Chomsky and Halle do not speak of syllables, but rather propose that /g/ falls before an /n/ which is followed by either a full or internal word boundary (i.e., ## or #).

Perhaps the best-known argument against a level intermediate between the systematic phonetic and systematic phonemic is presented by Halle (1959: 22–23) and reproduced in Chomsky (1964:100–101). The claim is made that recognizing a phonemic level will, in the words of Chomsky (1964:100), "destroy . . . the generality of rules, when the sound system has an assymetry." The example comes from Russian, which has the following phonological rule:

$$\underset{\mathbf{C}}{[-\text{son}]} \rightarrow [+\text{voice}] \,/\, \_ \begin{bmatrix} -\text{son} \\ +\text{voice} \end{bmatrix}$$

An obstruent becomes voiced before a voiced obstruent. Thus, a sequence of /t/ followed by /b/ will be pronounced [db], but a sequence of /t/ followed by /l/ will be pronounced [tl], since [l] is a sonorant. The problem Halle points out is that while there is a phonemic contrast between /t/ and /d/ in Russian, there is no contrast between the phoneme /č/ (which exists in Russian) and the phoneme /ǰ/ (which does not exist). And despite the fact that there is no voice contrast in the palatals, the same facts are observed with respect to the voicing rule. That is, a sequence of /č/ followed by /b/ will be pronounced [ǰb] (and, of course, /č/ followed by /l/ will remain [čl]). Since a strict phonemic analysis adhering to phonetic similarity (biuniqueness) would be forced to analyze [db] as /db/ (although the [d] represents a neutralization of /t/ and /d/ in Prague School terminology), the following rule is a *morphophonemic* rule:

$$\{\text{t}\} \rightarrow /\text{d}/ \quad / \_ \begin{bmatrix} -\text{son} \\ +\text{voice} \end{bmatrix}$$

That is, it changes a morphophoneme into a phoneme. The following rule, however, is a *phonemic* rule, since it merely states the allophonic distribution of the phoneme /č/:

$$/\text{č}/ \rightarrow [\text{ǰ}] \quad / \_ \begin{bmatrix} -\text{son} \\ +\text{voice} \end{bmatrix}$$

Thus, although these two rules are clearly instances of the same rule (as formalized in features above), they must be stated at different places in the grammar. Assuming both a morphophonemic and a phonemic level, the first rule converts a morphophonemic representation to a phonemic one and the second converts a phonemic representation to a phonetic one. In order to avoid this duplication (or lack of generality), it is necessary to reject the level of autonomous phonemics and recognize only a systematic phonemic level and a systematic phonetic level.

It would be unwise to suggest that all of Chomsky's (1964) criticisms apply to all schools of phonemics. The above argument is of course limited, since many phonemicists allowed neutralization of just the type found in Russian.

Thus it would appear that it is not so much a question of establishing a difference between a systematic phonemic level and a phonemic level, the first of which is valid and the second invalid, but rather a question of properly defining what the characteristics of the one valid *phonological* level are.

While it is clear that the phonological level can differ considerably from the phonetic representation, generative phonologists themselves are now debating the question of just how "abstract" phonology is. Probably most generative phonologists would agree that the words 'advice' and 'wheel' in German, both pronounced [ra:t], should be represented phonologically as /ra:t/ and /ra:d/, respectively (see Vennemann, 1968a). But representing [riyzayn] as /rē=sign/ is quite another story, for here we have to (1) represent the high front diphthongized vowel [iy] as abstract /ē/, (2) accept a special morphological boundary (=), and (3) represent [ay] as /ig/, that is, a radically different vowel with a consonant which is not realized phonetically (in this allomorph, at least).

There seem to be no constraints on the degree of abstractness allowable in generative phonology. For example, Lightner (1971) considers the possibility of taking the underlying forms of English back to a Proto-Germanic stage (before the application of Grimm's Law). He points out that there are alternations such as the following between [f] and [p], [ð] and [t], and [h] and [k]:

*f*oot	:	*p*edestrian
*f*ather	:	*p*aternal
*f*ull	:	*p*lenary
mo*th*er	:	ma*t*ernal
fa*th*er	:	pa*t*ernal
bro*th*er	:	fra*t*ernal
*h*eart	:	*c*ardiac
*h*orn	:	uni*c*orn
*h*ound	:	*c*anine

Perhaps the root of 'foot' should be recognized as the Latin-looking /ped-/? While almost no one would accept Lightner's proposal, his question is right to the point: "Where does one stop? And why?"

One way of trying to limit the powers of generative phonology is by looking at the nature of the rules that would be required. It is hard to imagine an environment for changing underlying /ped/ to [fʊt] other than by an arbitrary diacritic, for example, [+X]. The rule could then be written as follows:

$$p \rightarrow f / [+X]$$

But since there is no phonological or morphological correlate to this diacritic, this kind of rule would be equivalent to simply listing two forms in the lexicon, /fʊt/ and /pɒdɛstriən/.

Kiparsky (1968a) presented the first principled attempt to limit the powers of generative phonology. He distinguished between *contextual* and *absolute* neutralization. Contextual neutralization is the kind of situation we have seen in English (intervocalic /t/ and /d/ are neutralized), Fe?fe? (/k/ and /č/ are neutralized before /i/), and German (/t/ and /d/, among others, are neutralized syllable–finally). Typically, when there is a rule of the form,

$$A \rightarrow B / \_\_ C \quad \text{(that is, AC} \rightarrow \text{BC)}$$

and there are already [BC] sequences coming from another source, we say that /A/ and /B/ are neutralized before /C/. Absolute neutralization, on the other hand, occurs when there is a rule of the form

$$A \rightarrow B$$

and there are other instances of [B] coming from another source. The main difference between the two types of neutralization, then, is that in absolute neutralization the rule that accounts for the neutralization takes place without any context. That is, all instances of underlying /A/ merge with underlying /B/.

A concise example of absolute neutralization, which Kiparsky cites, comes from Sanskrit, which has the following CV sequences:

či	ku
ča	ka

Since there are no instances of **ki* or **ču*, /k/ and /č/ are in near complementary distribution—they contrast only before /a/. However, it would be possible to represent sequences of phonetic [ča] as underlying (systematic phonemic) /ke/, since there is no short [e] in Sanskrit, with the following derivations:

/ki/ → [či]
/ke/ → če → [ča]

The /k/ of /ke/ could be said to palatalize just like the /k/ of /ki/, yielding intermediate *če*. At this point a rule of the form

$$e \rightarrow a$$

would convert all instances of /e/ to [a], causing absolute neutralization with /a/.

Kiparsky argues that rules of this form, which create context-free neutralizations, should be disallowed, and he presents arguments from historical linguistics to support his position. Notice, first, however, that it is not the *form* of this rule of absolute neutralization that makes it so objectionable. This rule can in fact be rewritten with a context, as follows:

$$e \rightarrow a / č \_\_$$

In a sense this restatement is a trick, since it just so happens that all instances of underlying /e/ will occur after [č] at this stage in the derivation; /e/ is posited only after /k/ (which will in turn palatalize to [č]). The real objection seems to be simply calling something what it is not. That is, the argument should be stated as one against "imaginary" segments (Crothers, 1971).

One such imaginary segment is the /œ/ which Chomsky and Halle (1968) posit as the phonological representation of the English diphthong [ɔɪ]. While a rule of the form

œ → ɔɪ

does not involve neutralization (since there is no other source of [ɔɪ]), the postulated /œ/ of *boy* /bœ/ is at least as "abstract" as the underlying /e/ considered for Sanskrit.

This reinterpretation of the problem is visible in the Yawelmani case raised in the argument against Kiparsky by Kisseberth (1969). In Yawelmani, the following surface phonetic vowels are found:

i u
a ɔ e: a: ɔ:

Kisseberth argues that all instances of [e:] should be represented phonologically as /i:/, and some instances of [ɔ:] should be represented as /u:/, others as /ɔ:/. This would produce the more symmetric inventory of both long and short /i, a, u, ɔ/. His arguments are as follows.

First, there is a class of verbs of the underlying structure /CCV(C)/ which Kuroda (1967) terms "echo verbs." A phonological rule inserts a vowel between the first two consonants in the following way:

a CCe:(C) → CiCe:(C)
b CCa:(C) → CaCa:(C)
c CCɔ:(C) → CuCɔ:(C)
d CCɔ:(C) → CɔCɔ:(C)

Notice that cases **b** and **d** involve complete copying of the stem vowel, though the copied vowel is always short. Having noticed this, if we were to analyze verbs of class **a** as underlying /CCi:(C)/, then this /i:/ would also be copied as [i]. Similarly, if those verbs of the form [CuCɔ:(C)] were recognized as underlying /CCu:(C)/, then the copying rule would be completely general:

$$\emptyset \rightarrow \underset{[-\text{long}]}{V_i} \quad / \# \, C \_\_ C \, V_i, \qquad \text{where } V_i = V_i$$

A short version of the underlying vowel (V_i) of echo verbs is copied by this rule.

Another argument Kisseberth (1969) gives for his /i:/ and /u:/ solution comes from vowel harmony. While the aorist (past indefinite) suffix is represented phonologically as /hin/, it is converted to [hun] after some instances of phonetic [ɔ:]:

ćuyɔ:hun 'urinated'
hɔyɔ:hin 'named'

As seen from the copied vowel [u] in 'urinated,' this verb is represented phonologically as /ćyu:/. First the vowel /u:/ is copied to yield intermediate *ćuyu:*, and then the long vowel /u:/ is lowered to [ɔ:]. This solution ties in neatly with the vowel harmony occurring in the aorist suffix. It is just those verbs with underlying /u:/ which harmonize /hin/ to [hun]. That this is correct is seen from the fact that short /u/, but not short /ɔ/, also harmonizes /hin/ to [hun]:

hudhun 'recognized'
gɔphin 'took care of an infant'

Thus, /hin/ becomes [hun] after the stem vowels /u:/ and /u/.

This solution requires a rule of the following form:

a $\begin{bmatrix} i: \\ u: \end{bmatrix} \rightarrow \begin{bmatrix} e: \\ ɔ: \end{bmatrix}$

Notice that only part **b** of this rule involves absolute neutralization, since /u:/ and /ɔ:/ merge as [ɔ:] in all environments, while no merger occurs when /i:/ is converted to [e:]. Although only the latter part of the rule involves absolute neutralization, both involve setting up "imaginary" forms, that is, phonological forms which do not exist on the surface and which are converted to phonetic forms in a context-free fashion. The derivations for 'urinated' and 'named' are therefore as follows:

/ćyu: + hin/	/hyɔ: + hin/	
ćuyu:hin	hɔyɔ:hin	(by vowel copying)
ćuyu:hun		(by vowel harmony)
[ćuyɔ:hun]	[hɔyɔ:hin]	(by vowel lowering)

So-called "imaginary" phonological representations characterize, at least to some extent, probably all schools of phonology. Consider, for example, the following phonetic vowel system of Nupe:

i u ɪ ĩ
e o ɜ̃
a

Although Nupe has five phonetic oral vowels, it has only three phonetic nasalized vowels (Smith, 1967; Hyman, 1970b). The question, however, is how the vowel [ɜ̃] should be interpreted. Since there is an oral vowel /a/,

pattern considerations suggest representing this vowel as /ã/, the decision reached by Smith (1967). Since /ĩ/ and /ũ/ tend to be pronounced [ĩ] and [ũ], a low-level phonetic rule is postulated which changes all nasalized vowels to [−tense]. In a sense this amounts to recognizing an imaginary segment. While in this case the distance between the phonological and phonetic representations may seem negligible, no satisfactory way of measuring such "distances" has been proposed.

In the absence of theoretical constraints on abstractness, such as the one proposed by Kiparsky, a number of competing analyses will be possible of the data of many languages, for example, a very abstract analysis, a not-too-abstract analysis, a very nonabstract analysis. Since generative theory attempts to provide the one descriptively adequate grammar of a language, which is said to have psychological reality, proposals which limit the number of possible analyses for any given data represent claims about the nature of sound systems, which can in part be experimentally tested (see M. Ohala, 1974).

Since Kiparsky's unpublished paper, a number of papers, in addition to Kisseberth (1969), have defended certain "abstract" analyses. A final example of a possible abstract solution, again from Nupe, is presented in Hyman (1970a), where it is suggested that [C^{w}a] and [C^{y}a] should be represented, phonologically, as /Cɔ/ and /Cɛ/, respectively. Since consonants are normally labialized before /u/ and /o/, and since they are normally palatalized before /i/ and /e/, we can simply extend the labialization and palatalization processes to include /ɔ/ and /ɛ/, as seen in the following rules:

$$C \rightarrow C^{w} / \_ \begin{Bmatrix} u \\ o \\ ɔ \end{Bmatrix} \qquad \text{(LR)}$$

$$C \rightarrow C^{y} / \_ \begin{Bmatrix} i \\ e \\ ɛ \end{Bmatrix} \qquad \text{(PR)}$$

After /Cɔ/ and /Cɛ/ have undergone the labialization rule (LR) and the palatalization rule (PR), respectively, the following absolute neutralization (AN) rule applies:

$$\begin{Bmatrix} ɔ \\ ɛ \end{Bmatrix} \rightarrow a$$

Since /ɔ/ and /ɛ/ neutralize in a context-free fashion with /a/, this is a case of absolute neutralization, as defined by Kiparsky (1968). We can, however, provide a context for this rule, as follows:

$$\begin{bmatrix} ɔ \\ ɛ \end{bmatrix} \rightarrow a / \begin{bmatrix} C^{w} \\ C^{y} \end{bmatrix} \_$$

The rule now states that /ɔ/ becomes [a] after [C^{w}] and /ɛ/ becomes [a] after [C^{y}]. This rule not only directly incorporates the motivation for the rule (that is, the fact that the labiality and palatality of /ɔ/ and /ɛ/ have been transferred to the consonant), but also captures the fact that each instance of [a] can be easily identified as deriving from /ɔ/, /ɛ/, or /a/ on the basis of the preceding consonant, as seen in the following derivations:

/tɔ́/	→ t^{w}ɔ́ →	[t^{w}á]	'to trim'
/tɛ́/	→ t^{y}ɛ́ →	[t^{y}á]	'to be mild'
/tá/	→	[tá]	'to tell'

Two kinds of evidence for this /ɔ/ and /ɛ/ solution were proposed. First, it was claimed that reduplication provides evidence for the underlying vowel /ɔ/. As seen in the following examples,

tí	'to screech' →	tītí	'screeching'
tē	'to break' →	tītē	'breaking'
tá	'to tell' →	tītá	'telling'
tú	'to ride' →	tūtú	'riding'
tò	'to loosen' →	tūtò	'loosening'

the reduplicated vowel is [i] if the stem vowel is [−round], that is, /i/, /e/, or /a/; or [u] if the stem vowel is [+round], that is, /u/ or /o/. Notice, however, the following forms:

t^{w}á	'to trim' →	tūt^{w}á	'trimming'
t^{y}á	'to be mild' →	tīt^{y}á	'being mild'

The expected form for 'trimming,' if /t^{w}/ is taken to be an underlying consonant, is **t^{w}ītʷá*. If, on the other hand, we recognize the underlying form /tɔ́/, this /ɔ/ naturally falls into the same class with /u/ and /o/, and the automatically chosen reduplicated vowel is [u].

The second argument is based on the findings of Hyman (1970b) concerning the nature of foreign sound assimilations in borrowing. It was argued in Hyman (1970a) that since Yoruba [Cɔ] and [Cɛ] come into Nupe as [C^{w}a] and [C^{y}a], the rule of absolute neutralization must be considered productive. Some relevant examples are:

Yor.	[kɛ̀kɛ̌] >	Nupe	[k^{y}àk^{y}á]	'bicycle'
Yor.	[ɛ̀gbɛ̀] >	Nupe	[ègb^{y}à]	(a Yoruba town)
Yor.	[tɔ̄rɛ̄] >	Nupe	[t^{w}ār^{y}ā]	'to give a gift'
Yor.	[kɔ́bɔ̂] >	Nupe	[k^{w}áb^{w}à]	'penny'

According to this argument, the rule of absolute neutralization is responsible for these borrowings, and for the fact that Nupes, when they speak Yoruba, frequently replace Yoruba [Cɔ] and [Cɛ] with Nupe [C^{w}a] and [C^{y}a]. For justification of this kind of argumentation see Hyman (1970b) (and also Ohso, 1971, and Lovins, 1973, for more recent work on this subject).

The question of how Nupe should be analyzed has been raised a number of times since the original abstract solution was proposed (see, for instance, Harms, 1973, and, for a reply, Hyman, 1973d; also Crothers, 1971; Vennemann, 1973; Krohn, 1974). Just how abstract phonology is remains a question that has yet to be answered in a manner satisfactory to all.

3.4 General Considerations in Setting Up Underlying Forms

In preceding sections we saw basically three approaches to phonological analysis, which can be summarized here by means of the following example from English. As seen in the following forms,

im–possible
in–determinate
iŋ–congruous

the prefix meaning 'not' is pronounced [ɪm] before labials, [ɪn] before alveolars, and (at least optionally) [ɪŋ] before velars. The question is, how should these forms be represented phonologically? In a strict phonemic approach one might argue that the phonetic and phonological representations are identical, that is, that these prefixes should be analyzed as the allomorphs /ɪm/, /ɪn/, and /ɪŋ/, respectively. Such phonologists would point out that since the words *ram* [ræm], *ran* [ræn], and *rang* [ræŋ] show a three-way nasal contrast, the phonemes /m/, /n/, and /ŋ/ are required in English. It should be recalled that in this first view the phoneme was defined as a class of sounds having phonetic similarity (see **3.1**). Thus, by the principle of "biuniqueness" (see **3.2.1**), the sounds [m], [n], and [ŋ] are assigned to the phonemes /m/, /n/, and /ŋ/ of the negative prefix, just as they are in the case of *ram*, *ran*, and *rang*.

A second solution invokes the notion of neutralization from Prague School phonology. Since nasals do not contrast before such consonants, this morpheme can be represented as /ɪN–/, that is, with an archiphoneme nasal which is specified as [+cons, +nasal], but which is left unspecified for place of articulation. This solution then captures an important fact missed by the strictly phonemic solution, since it recognizes /m/ and /n/ only where these two phonemes contrast, and recognizes /N/ where there is no contrast.

A weakness of both these solutions, however, is the fact that when this prefix is followed by a vowel, its realization is [n]. If one were to start with underlying /m/, /n/, and /ŋ/, then there would be no way to capture the fact that the basic or unassimilated form of this prefix is [ɪn], as in the word *inability*. The same problem is inherent in the archiphoneme approach. On

the other hand, if one were to start with the representation /ɪn/, a rule of homorganic nasal assimilation, written as follows,

$$n \rightarrow [\alpha \text{ place}] / \underset{C}{\_} [\alpha \text{ place}]$$

would state that /n/ assimilates to the place of articulation of the following consonant. Thus, underlying /ɪn/ is realized as [ɪm] before labial consonants (*im-possible*) and as [ɪŋ] before velar consonants (*in-congruous*). Before alveolar consonants and before vowels (*in-determinate* and *in-ability*), it is realized as [ɪn].

Setting up one basic underlying form from which predictable allomorphs or alternations can be derived runs into some difficulty, however, since, as pointed out above, there seems to be no constraint as to how "abstract" the base form can be. For example, while there is a productive rule of homorganic nasal assimilation of the type seen above, we are faced with the problem of what to do with words such as *illegal* and *irregular*, where the assimilation of the /n/ of this same negative morpheme is complete. That is, /n/ assimilates to [l] before [l] and [r] before [r], and presumably the resulting [ll] and [rr] sequences are later simplified to [l] and [r], respectively. Are the underlying representations *in-legal* and *in-regular* too distant from the phonetic representations? While phonologists disagree about the permitted degree of abstractness, all those working in the framework of generative phonology accept the notion of a base "underlying form" from which allomorphs are derived by phonological rules. With this in mind, we can now ask, what are the general considerations in determining underlying forms?

3.4.1 Predictability

Given a phonological alternation, such as the alternation between [t] and [d] in the German words *Rad* [ra:t] 'wheel' and *Räder* [rɛ:dər] 'wheels,' how does one decide which of the two phonetic realizations is closest to the underlying representation? Or, in other words, how does one determine the "basic allophone"? While there is no foolproof rule or "discovery procedure," there are some general criteria which are sometimes cited by phonologists. The first criterion is *predictability*. Often there is little cause for hesitation, since the various alternations can be phonologically predicted (that is, by rule) only if one starts with one of the allophones—but could not be predicted if one started with the other. The German case of final devoicing is an example. If the word 'wheel' is represented with a /d/ underlyingly, that is, /ra:d/, then a rule of final devoicing would change /d/ to [t] in [ra:t], but not in the plural form [rɛ:dər]. The rule that converts /b, d, g, v, z/ to [p, t, k, f, s] can be written as follows:

$$[-\text{son}] \rightarrow [-\text{voice}] / \_\ \$$$

Voiced obstruents are devoiced in syllable-final position. If, on the other hand, 'wheel' were to be represented with underlying /t/, that is, /ra:t/, then a rule would be required which would convert /p, t, k, f, s/ to [b, d, g, v, z] in some environment, so that /ra:t/ + /¨ər/ (where ¨ represents the umlauting process that fronts [a:] to [ɛ:]) is realized as [rɛ:dər] and not as *[rɛ:tər]. However, notice that the plural of Rat [ra:t] 'advice' is *Räte* [rɛ:tə]. Since both 'wheel' and 'advice' would presumably be recognized as /ra:t/ in this analysis, there would be no way of predicting which cases of final /t/ become [d] and which remain [t]. Since we can predict the alternations in one direction only, we assume that 'wheel' should be represented phonologically as /ra:d/ and that there is a rule of final devoicing.

Of course, it would be possible to maintain both 'wheel' and 'advice' as /ra:t/ if we used some arbitrary diacritic mark, say [+D], to identify those morphemes whose final /t/ becomes [d] by rule. By using such diacritics, the claim is made that this is not a purely phonological alternation, but rather a partly morphological one, since morphemes must be identified. Phonologists have generally argued that diacritics, while necessary to capture irregularities in languages, represent complexities and should be used only when strictly phonological solutions (that is, ones using distinctive features only) cannot be motivated. Since the German rule can be written in strictly phonological terms, the use of diacritics is ruled out.

A second example of the predictability criterion comes from Maori (Hale, 1971, as reported in Kiparsky, 1971). In Maori there is an alternation between certain consonants and Ø (that is, zero), as seen in the following examples:

VERB	PASSIVE	GERUND	GLOSS
hopu	hopukia	hopukaŋa	'to catch'
aru	arumia	arumaŋa	'to follow'
tohu	tohuŋia	tohuŋaŋa	'to point out'
maatu	maaturia	maaturaŋa	'to know'

As seen in the leftmost column, the active form of these verbs ends in a vowel, in this case [u]. In the passive and gerund forms, however, different consonants appear on the surface, in this case [k, m, ŋ, r]. There are two possible solutions. First, one might set up underlying forms which end in consonants. In this case we would recognize the underlying forms /hopuk/, /arum/, /tohuŋ/, and /maatur/, and a rule which deletes word-final consonants:

$$C \rightarrow \emptyset \,/\, \_\#\#$$

The second solution recognizes the underlying forms /hopu/, /aru/, /tohu/, and /maatu/, and a rule of consonant insertion. However, in this case there is a problem in predicting the exact identity of the consonant which will appear. There is no reason in this solution why /hopu/ should take a [k] but /aru/ should take an [m]. In other words, we are again forced into marking such forms with diacritics, for example, [+K], [+M], etc. Since Ø can be predicted from underlying final /k, m, ŋ, r/, but since [k, m, ŋ, r] cannot be phonologically predicted from Ø, the first solution is preferred. Notice also that there are some cases of verb forms ending in [u] which do not take *any* consonant, for example, [patu] 'to strike,' passive [patua], gerund [patuŋa]. (The expected passive [patuia] and gerund [patuaŋa] are simplified by rule.) This verb will therefore be represented as /patu/. (For more discussion of this Maori data, see **5.2.8.**)

3.4.2 Economy

In phonemic analysis, a solution is judged to be more economical than another if it recognizes fewer phonemes. While this notion has not been explicitly incorporated into generative phonology, it is sometimes invoked in terms of overall "simplicity" (see **4.1**) by generative phonologists. One example is English *ng*. A solution recognizing a word such as *sing* as /sɪŋ/ is forced to admit an additional phoneme. A solution representing this word as /sɪng/, since it avoids a phoneme /ŋ/, is more economical. However, economy in the number of phonemes or underlying segments frequently entails a greater complexity in the phonological rules. As seen in **3.3.1**, if we recognize /sɪng/ we need to apply a rule of homorganic nasal assimilation (which we already know characterizes English—compare /ɪn–/), which yields the intermediate form [sɪŋg]. At this point we need to introduce a rule *not* previously needed, namely, one which deletes the [g] of *sing*, thereby giving the phonetic form [sɪŋ]. Notice that neither solution can be argued for by the criterion of predictability. If we recognize an underlying /ŋ/, then a [g] will have to be inserted into the word *longer* [lɔŋgər] (compare *long* [lɔŋ]), but not in the word *singer* [sɪŋər]. If we recognize only /ng/, then the /g/ will have to be deleted in *singer*, but not in *longer*. Thus, both solutions require nonphonological information, namely boundary information. As proposed by Chomsky and Halle (1968:85n), the underlying forms of *longer* and *singer* are recognized with different internal grammatical boundaries, /sɪng#ər/ and /lɔng+ər/. Post-nasal /g/ is deleted before a word boundary (#), as in *sing* and *singer*, but not when there is only a morpheme boundary (+), as in *longer*, or no boundary, as in *finger* [fɪŋgər] (see **6.1.2.2.**).

3.4.3 Pattern Congruity

This criterion was cited by certain American phonemicists (for example, Swadesh, 1934:36), who saw the phoneme as a (psychological) point in a *pattern* (compare Sapir, 1925). In this view, a solution can be argued for on the basis that it conforms to the overall pattern of the phonological system. The /ng/ solution is a good example. If a separate phoneme /ŋ/ were

recognized, we would have to ask why it, unlike /m/ and /n/, cannot appear at the beginning of a word. If, on the other hand, /ng/ is posited, the failure of [ŋ] to appear at the beginning of words in English can be explained by reference to a more general overall pattern; namely, just as /mb/ and /nd/ sequences do not occur initially, neither does /ng/ (whose phonetic reflex is sometimes [ŋ]).

The use of pattern congruity as a criterion has led many phonologists to seek segments to fill "holes" in the pattern. For example, the following consonants represent the phonetic consonant system in Fe?fe?–Bamileke (ignoring aspirated consonants):

p	t	č	k	?
b	d	ǰ	g	
f	s	š		h
v	z	ž	ɣ	
m	n	ɲ	ŋ	
	l			
w		y		

The columns represent places of articulation, the rows manners of articulation (respectively, voiceless stops, voiced stops, voiceless fricatives, voiced fricatives, nasal consonants, liquids, and glides). A number of holes in the pattern are observed in the above chart. In addition, a number of consonants stand by themselves (for example, [l]). Thus, typically, the consonants which are isolated are frequently moved into positions which are vacant in the more general pattern. For example, Fe?fe? has no voiceless velar fricative [x]. It does, however, have a glottal fricative [h], which we can conveniently move into the velar slot to complete the series. Other rearrangements can be effected to yield the following phonetic chart:

p	t	č	k
b	d	ǰ	g
f	s	š	h
v	z	ž	ɣ
m	n	ɲ	ŋ
w	l	y	?

Other movements are the following: since the glides [w] and [y] are made at a different point of articulation from [l], the two series are collapsed; since there is no back glide, the glottal stop has been moved into that position. Notice that the bottom row contains segments which Chomsky and Halle (1968) regard as [+son], though the case for treating a glottal stop as a sonorant is weak. While the consonant system has been made to look symmetric, this has been at the expense of calling some phonetic segments something they are not—for example, [?] is not a sonorant, [h] is not velar. While by Sapir, who viewed phonemic structure as points in a pattern, such

arrays of sounds as seen above were accorded theoretical status, to other phonologists such patterns merely summarize the phonetic segments of a language. Thus, as reported in Hyman (1972b), the underlying (systematic) phonemes of Fe?fe? are as follows:

	t	č	k	?
b	d	ǰ	g	
f	s			h
v	z			
m	n		ŋ	
(w)				

(The /w/ is of questionable status.) Thus, phonemically, a number of holes do exist in the pattern.

This manipulation is most frequently observed, perhaps, in the way phonologists present vowel systems. In vowel systems with the three vowels /i, u, a/, the five vowels /i, e, u, o, a/, or the seven vowels /i, e, ɛ, u, o, ɔ, a/, /a/ is often represented as a low *central* vowel, thereby giving the impression of symmetry:

i		u	i		u	i	u
	a			e	o	e	o
				a		ɛ	ɔ
							a

In vowel systems with the four vowels /i, u, o, a/, the chart is usually presented as

i	u	rather than	i	u
a	o			o
			a	

even though /a/ is lower in vowel height than /o/ and is not necessarily a front vowel. In this case, however, the symmetric vowel chart captures the the fact that in such languages there is phonologically only a two-way vowel height contrast and a two-way front/backness contrast. But to be consistent, three-vowel systems should be written as in **a** or **b**:

a	i	u	**b**	i	u	
		a			a	

Such diagrams represent the two possibilities for the phonological patterning of /i/, /u/, /a/: in **a** /u/ and /a/ pattern together, as opposed to /i/, since they are both [+back]; in **b** /i/ and /a/ pattern together, since they are both [−round]. In the first language we should expect /u/ and /a/ to function together in phonological rules, while in the second language we should expect /i/ and /a/ to function together.

One of the most frequent references to pattern congruity in phonemic analysis concerns the question of whether something should be analyzed as one phoneme or two. For instance, in a language with an aspiration contrast, such as Thai, one might ask whether the contrast should be represented as /p/ vs. /pʰ/ or as /p/ vs. /ph/. In the case of palatalization, one might wonder whether to set up a series of palatalized consonants (for example, /pʸ/) or a two-phoneme sequence of consonant followed by /y/ (for example, /py/). Such questions can frequently not be answered by the phonetics alone, but only by referring to the overall pattern of the language—in particular, the general canonical shape of syllables. In Igbo, for instance, syllables generally consist of a single consonant followed by a single vowel (that is, CV). The major exception to this pattern is the presence of labialized velars, which could possibly be analyzed as /kw/, /gw/, and /ŋw/. However, if they were to be analyzed as /kʷ/, /gʷ/, and /ŋʷ/, that is, as single consonants with a secondary articulation, then they would not violate the syllable structure of the language. If, on the other hand, we were to accept the two-phoneme analysis, then the system would be broken, and we would have no explanation of why /w/ only occurs after /k/, /g/, and /ŋ/. In the one-phoneme solution we simply say that the language has labialized velars, and, since labialized velars are much more frequent and expected in languages than labialized labials or labialized dentals, no further statement is required.

Another consideration in deciding whether to derive a given phone or phones from one or two phonemes is whether the individual components are found in isolation in the language. For instance, we could not analyze aspirated stops as /ph/, /th/, and /kh/ in a language where /h/ does not appear alone. Similarly, the phonological representations /py/ and /pw/ would be avoided in languages that do not exhibit /y/ and /w/ functioning as independent consonants. This consideration is an extension of what is known in European phonology as the *commutation test* (Fischer-Jørgensen, 1956; Martinet, 1960:73). From a minimal pair such as *lamp* and *ramp* in English we conclude that there is a distinctive contrast between the two phonemes /l/ and /r/. Now, from a minimal pair such as *ramp* and *cramp*, we conclude that there is a distinctive contrast between Ø and /k/, and that *cramp* must therefore be analyzed as having an initial consonant cluster, rather than a single initial consonant. Finally, the minimal pair *ramp* and *amp* shows that *ramp* must be analyzed as having four phonological units, since /r/ contrasts with Ø (compare *camp* and *amp*). Martinet (1960:74) applies this test to the English *ch* sound. The question is whether this should be analyzed as /č/ or /tš/, that is, as one phoneme or two. He points out that English has not only the word *chip* [tšɪp], but also the word *ship* [šɪp]. From this opposition of *č*:*š* (where *č* = *tš* phonetically), we conclude that the [t] of [tšɪp] contrasts with Ø. From the opposition between *chip* and *tip* [tɪp], we conclude that the [š] of [tšɪp] contrasts with Ø. Therefore, *chip* should be analyzed by this criterion as /tšɪp/. On the other hand, since Spanish has this alveopalatal affricate (for example, *mucho* 'very') but does not have the corresponding fricative [š], *mucho* must be analyzed as /mučo/.

While the commutation test yields these results, Martinet rightly rejects the two phoneme /tš/ for English. He again appeals to the notion of pattern congruity. He points out that this [tš] sound must be analyzed exactly as the corresponding voiced *j* [dž] in English. Now, while there is a word *gyp* [džɪp] and a word *dip* [dɪp], there is no word *[žɪp] in the language. In other words, [ž] must always be preceded by [d] when it occurs at the beginning of a word. Since this is the case, [dž] must be analyzed as one phonological unit, that is, as /j/. And since Martinet wants to analyze the *ch* sound in like fashion, he argues that the first argument from commutation should be given up in favor of the pattern, and so we recognize underlying /č/. (For more on the question of one vs. two phonemes, see **4.4.1.**)

This, of course, points out the arbitrariness of this criterion, since it is possible that each of two conflicting analyses breaks the pattern in a different way. One wonders, for example, why /j/ should not be reanalyzed as /dž/, on analogy with /tš/, and not vice-versa. Notice, finally, that patterns change through time. The Grebo language (Innes, 1966) generally exhibits a CVCV pattern, but it has begun to syncopate vowels in fast speech (for example, /fodo/ 'emptiness' becomes [flo] in rapid speech), such that there are now syllables of the form CLV. With time we can expect the CLV forms to take precedence over and eventually drive out the CVCV forms. In fact, there are some forms, mostly borrowed, which only exist in their CLV form, for example, [fli] 'flea.' Thus, whenever an argument is made for conforming to a pattern, for example, CVCV, we have to be sure that the language is not on the way to establishing another pattern. It may be that the old pattern is no longer the criterion for congruity.

3.4.4 Plausibility

A fourth criterion that is often invoked is *plausibility*. Given two possible solutions, is there one which in some sense is more plausible (or "natural"—see Chapter 5)? Consider, for example, a language which has the following phonetic sequences (Nupe comes close, although it also has [ša]):

ši	su
še	so
sa	

The alveopalatal fricative [š] is found before [i] and [e] and the alveolar

fricative [s] before [u], [o], and [a]. Thus, we have a classic case of complementary distribution. There are two possible solutions. First, we can recognize underlying /si, se, su, so, sa/ and posit a rule such as

$$s \rightarrow š \,/\, \_\_ \begin{Bmatrix} i \\ e \end{Bmatrix}$$

which converts /si/ and /se/ to [ši] and [še], respectively. Or we can recognize underlying /ši, še, šu, šo, ša/ and posit a rule such as

$$š \rightarrow s \,/\, \_\_ \begin{Bmatrix} u \\ o \\ a \end{Bmatrix}$$

which converts /šu/, /šo/, and /ša/ to [su], [so], and [sa], respectively. The first solution is plausible, while the second solution is implausible. Recognizing only /s/ is plausible, because the rule which derives [š] before /i/ and /e/ is a natural assimilation rule. That is, when /si/ becomes [ši], the alveolar /s/ assimilates to the frontness (or palatality) of /i/. Similarly, when /se/ becomes [še] the same assimilatory process is observed. On the other hand, if we start with underlying /š/, the rule which is required to derive [s] before /u/, /o/, and /a/ is not a natural assimilation rule. While the process of a palatal consonant becoming nonpalatal before a nonpalatal vowel would appear to be assimilatory in nature, the question is why /š/ should become more fronted (that is, to [s]) rather than backed (to, say, [x]) before the back vowels in question. Thus, this rule seems to be unmotivated from a phonetic point of view.

Rule plausibility usually refers to *phonetic naturalness*. Certain phonological rules are found to occur frequently in languages, and the reason for this frequency is the fact that segments tend to assimilate to neighboring segments, and they do so in fairly predictable ways (see Schachter, 1969; Schane, 1972). The notion which is usually brought forth to explain these phenomena is *ease of articulation*. It is claimed to be easier to pronounce [ši] than [si], since in the first case both segments agree in palatality.

What this means is that plausible phonological rules are usually unidirectional. Thus, one can use this criterion in phonological analysis and try to establish an inventory of underlying segments from which the surface segments can be derived by plausible rules. This criterion, like the other criteria, is subject to other considerations. In particular, some languages do have implausible or "crazy" rules (Bach and Harms, 1972). As discussed in **5.2.3**, the most phonetically natural rule is not necessarily the most simple rule. However, as a general principle, plausibility or rule naturalness is an important criterion in conducting phonological analyses.

4

PHONOLOGICAL SIMPLICITY

4.1 Simplicity, Economy, and Generality

In **3.4.2**, the notion of economy was said to be one of the criteria often used as a guide in phonemic analysis. A solution with fewer phonemes is judged more economical than a solution recognizing more phonemes. Similarly, we might say that a solution using fewer rules is more economical than a solution requiring more rules, and so on. Economy, then, is a quantitative measure by which a given solution can be evaluated as requiring fewer or more mechanisms (phonemes, rules, conventions, etc.) than another solution. This notion is characteristic of phonemic approaches to phonology, and, as we shall see, has its application in the history of generative phonology as well.

While one might be tempted to view a solution recognizing fewer phonemes as "simpler" than a solution recognizing more phonemes, there is another view which equates simplicity with generality. In terms of the phonemic inventory, the following argument might be made:

S_1			S_2		
i	ü	u	i	ü	u
e		o	e	ø	o
ɛ		ɔ	ɛ	œ	ɔ
	a			a	