Phonology Yearbook 3 (1986) 3–26
Printed in Great Britain

*Consumer's guide to evidence in phonology**

John J. Ohala
University of California, Berkeley

ABSTRACT

The purpose of evidence adduced in support of a theory is not to prove that theory true but to demonstrate that competing theories account for facts less well and thus no longer demand our attention. Evidence therefore, if it can unambiguously decide between competing theories, helps the discipline to spend its resources on only a few issues at a time. I evaluate this winnowing capacity of various kinds of evidence which have been offered in support of hypotheses on what knowledge native speakers have about the sound patterns in their language and how they use it: surface sound patterns, sound change, poetry, speech errors, word games, and experiments. I argue that experiments provide evidence of the highest quality. Four experiments are reported, three psychological and one phonetic, which offer evidence on the following claims: (a) the psychological basis of speakers' awareness of phonotactics, (b) speakers' awareness of the morphemic constituents of complex derived words, (c) whether epenthetic stops of the sort evident in words like *team[p]ster* are added by purely mechanical constraints of the articulatory apparatus or whether they can be attributed to higher, pre-phonetic levels, and (d) the factors that determine speakers' assignment of allophones to phonemes.

1 Introduction

For the past 30 years phonologists have speculated on how sound patterns in language are represented in the human mind (Chomsky & Halle 1968: viii). The claims made, of course, are only as good as the evidence they are based on. Accordingly, there is often debate over the relative merits and relevance of evidence of various sorts, especially 'internal' evidence as opposed to 'external' evidence (see, e.g., Saussure 1916: ch. 5; the summary report of the symposium on 'Phonology' in the *Proceedings of the 9th International Congress of Phonetic Sciences*, Vol. 3 (1980), Copenhagen, pp. 59–75). Internal evidence seems to be based on any of the kinds of data that can be gleaned from a dictionary, grammar, text, or a transcription of speech, e.g. allophonic variation, constraints on sound sequences, word sandhi, dialect variation, morphophonemic variation. External evidence, on the other hand, comes from non-traditional sources:

word games, speech errors, mistakes made by first and second language learners, experiments of any kind (phonetic, psycholinguistic, sociolinguistic), etc.

This is a very curious distinction to make. There seems to be no single relevant property inherent to these sources of evidence that would allow one to classify them as internal or external. It cannot be a matter of the data collection being done *in vivo* (i.e. under natural conditions) as opposed to *in vitro* (i.e. under artificial circumstances): the field worker's elicitation of the pronunciation of words is typically done in as unnatural a language-using situation as are most psycholinguistic experiments. On the other hand, speech errors and texts can both be produced under natural conditions.

Apparently, however, the most important differentiating characteristic between them is that internal evidence comes from the kind of data that the majority of phonologists have been working with since the early nineteenth century whereas external evidence doesn't. Internal evidence is traditional and external evidence is 'new-fangled'.

Applying such a reactionary criterion to differentiate types of evidence is not only useless, I would maintain, it is also potentially damaging to the field if it leads anyone to disregard evidence from any source. Let us instead evaluate the types of evidence on the basis of how helpful they are to us as purveyors and consumers of phonological arguments.

We should first remind ourselves that the function of evidence is not to lead us to the truth. In addition to Popper's (1959) arguments on this point there is the testimony of history of science. Hypotheses are made, evidence is presented for and against them, some hypotheses are accepted and gain a following, but sooner or later they get supplanted by new hypotheses which have better evidence. There is no reason to expect that current claims will escape the same fate. What evidence does accomplish is to help us to choose between competing hypotheses. Rather than prove one hypothesis, it disproves competitors. This may appear at first to be a rather meagre accomplishment but upon further reflection its importance becomes clear. Naturally, the number of hypotheses that can be proposed as explanations for a given phenomenon are unlimited. If all the hypotheses which have accumulated throughout the history of a discipline still had to be seriously entertained by anyone working on a given problem, still had to be taught, learned, etc., this would be a serious drain on our time and energy. By eliminating non-competitive hypotheses we can focus our attention on and give temporary allegiance to one or at most just a few hypotheses at a time. A good measure of the quality of evidence commonly used by a scholarly discipline is the extent to which old hypotheses or theories do or do not continue to encumber it.

Although there is no way to achieve absolute truth, the evaluation methods used by the mature sciences have quite obviously led to *convergence*, where, typically, newer theories subsume the replaced theories or represent refinements of them. The history of disciplines that do not have

effective means of evaluating their claims often resembles a Brownian movement through the space of possible theories.

Compare, for example, physics and theology. The modern physicist interested in understanding the structure of matter is no longer expected to review the theories of the ancient Greeks on the topic, e.g. the hypothesis that all matter consists of earth, air, fire and water in various combinations. (This is not to deny, of course, that such theories are of keen interest to the philosopher and historian of science.) Evidence has been presented which enables us to discard those theories. In theology, on the other hand, it seems that the only way one of the many competing theories (religions) might be eliminated from the running is the death or enforced apostasy of its adherents. Even then, the demise of the theory may be only temporary. Druidism, I gather, is currently experiencing a revival.

I leave it to the reader to say whether linguistics is closer to physics or to theology in the way of giving out-dated theories a decent burial.

In any case, a serious evaluation of phonology's sources of evidence can only benefit the field. Every dollar, every hour, every gram of creativity that we spend now trying to raise the quality of evidence we use will eventually save us 100 times that in fruitless debate and wasted research efforts, not to mention an ocean of ink and forests of paper.

2 A comparison of different sources of evidence in phonology

I propose to evaluate *briefly* and in an *admittedly subjective* way some of the diverse sources of evidence used in phonology to support claims regarding the psychological mechanisms serving language use. It is primarily the 'resolving power' of evidence, i.e. its ability to help us choose unambiguously between competing hypotheses, which is evaluated. Contributing to this is the possibility of going back repeatedly to the same basic source and retrieving more useful evidence. This is important because any given piece of evidence is rarely definitive; inevitably another 'challenger' hypothesis will eventually appear which is as compatible with the known evidence as was the survivor of the last contest between hypotheses. To decide the new contest, one must seek evidence anew. It is important, then, to know whether some sources are exhaustible or inexhaustible suppliers of evidence.

2.1 Surface sound patterns

Ever since Saussure linguists have regarded the surface pattern in language as the reflection, perhaps a blurred one, of some underlying structure. Today, those trained in the generative tradition examine these surface

patterns for what they may reveal about underlying *mental* structures. For example, noticing such English word pairs as those in (1):

(1) Peter / petrify extreme / extremity
chaste / chastity profound / profundity
reside / residual grade / gradual

Chomsky & Halle (1968) assert that native speakers are aware of a systematic phonological relationship between the words in each pair and that their knowledge takes the form of (a) a common underlying form for both members of a pair (minus affixes), and (b) one or more rules which derive the phonetic forms from the underlying ones. Considerations of simplicity guide the linguist in finding these constructs.

But simplicity by itself has never been of much use in helping us choose between competing hypotheses. It is true that when empirical evidence is lacking we may give tentative allegiance (i.e. favourable consideration) to the hypothesis which seems simplest to us, but evidence must eventually be obtained to decide the issue. The problem, simply, is that simplicity has never been successfully defined and even if it were, there would be no guarantee that the universe (including speakers' brains) shares our conception of it and bias for it. Kepler tried in vain to establish 'simple' circles as the basis for planetary orbits and was greatly disappointed that his calculations pointed to their being ellipses.

Actually, even on grounds of perceived simplicity, another hypothesis immediately suggests itself about how native speakers accommodate pairs such as those in (1):

(2) They know each word as a separate item and although they may recognise a relationship between them, this relationship could as well be based on their semantic or orthographic properties as on their phonological aspects (Derwing & Baker 1977; Jaeger 1984, this volume; Wang 1985; Ohala & Ohala 1986; Wang & Derwing this volume).

Needless to say, the existence of such phonetic differences between related words is not a fact that needs a psychological explanation; they arose due to sound change. There is no more reason to expect the native speaker to be aware of the historical factors which gave rise to the phonological aspects of these alternations than there is to expect such knowledge in the case of the cognate pairs in (3) (see also Kahn 1976: 3ff):

(3) quick / biology queen / gynaecology
lobby / lodge snare / nerve
rabies / rage work / energy

We must conclude that surface patterns such as those in (1) offer us little help in deciding between these competing hypotheses. I do not say their value is zero because at the very least such evidence does allow us to discard hypotheses which are inconsistent with the given data. Thus, for the patterns in (1) we do not need to consider the hypothesis that speakers

know a rule that tenses a vowel in the environment —CVCV. This is a rather meagre accomplishment, though, because it still leaves us unable to decide between the large number of hypotheses that are consistent with the data.

The possibilities of sifting through surface sound patterns again and again for new evidence are quite limited if one views the lexicon as a fixed corpus. Only if the original data gathered were incomplete would a new search be likely to pay off. Finding pairs such as those in (4), where there is no alternation in the quality of the stressed vowels, could, if they had not been originally cited, disconfirm a hypothesis that the pattern in (1) was without exception:

(4) obese / obesity crude / crudity
probe / probity nude / nudity
immune / immunity note / notify
between / betweenity prose / prosify

If one views the corpus as open-ended – since speakers do create new words – then collections of previously unobserved surface patterns may help to decide between competing hypotheses. Naturally occurring neologisms such as *mundaneity* (< *mundane*, pronounced [mʌn'dejnɪɾi] and not *[mʌn'dænɪɾi], T. Armbruster, personal communication), or *musicism* (< *music*, pronounced ['mjuzɪkɪzm̩] and not *['mjuzɪsɪzm̩], L. Hyman, personal communication), could help to disconfirm a hypothesis that the pattern in (1) and 'velar softening' are fully productive. But this is like harvesting apples by sitting under an apple tree and waiting for the apples to drop. One might have to wait a long time.

Therefore, on a scale of 10, I rate the evidence from surface patterns at 1·0.

2.2 Sound change

Kiparsky (1968) cites data from sound change as support for the psychological reality of posited phonological constructs. If we could be certain that all sound change is a mentalistic phenomenon, i.e. due to language users choosing to change their grammars (subconsciously perhaps), then this would be appropriate. But this is not the case. Non-mentalistic mechanisms for sound change have been proposed for well over a century (Sweet 1874) and many of them are supported by laboratory evidence (Ohala 1974, 1981a, 1983a, b). Therefore, appealing to sound change to demonstrate the psychological reality of posited phonological constructs begs the question unless we can identify those aspects of sound changes which are motivated by mentalistic factors and those which are not. For example, the existence of sound patterns which are amenable to an analysis with ordered rules is least in need of a psychological account: they come about due to sound changes which are themselves ordered in time and which often interact, for example, when earlier changes either create or deplete forms subject to later sound changes. However, paradigm regular-

isation is a good candidate for psychologically motivated change and, furthermore, this has some empirical support (Bybee & Slobin 1982).

I therefore rate the evidence from sound change 2·0, a low score due to the extreme difficulty of making controlled observations of it.

2.3 Poetry

A song made popular by the singing group Peter, Paul, and Mary (and later John Denver) contains the lines in (5). Here *plane* [pʰlejn] is made to rhyme with *again* [əˈgɛn]:

(5) I'm leaving on a jet plane,
don't know when I'll be back again.

Phonetically the vowels in these two words obviously do not constitute a good rhyme (as pronounced by the singers). Could this, however, be used as evidence for these vowels being identical in their underlying (psychological) form? It might be if we could be sure of the source of knowledge the composer drew on when writing these lines. If established English-language conventions of rhyming prompted this choice of words, then the data contribute no insight into their underlying forms. (Regarding the quality of the second vowel in *again*, see Mencken 1948: 75–76.) Poetic conventions regarding permissible rhymes, metrics, etc., are usually quite conservative and once established tend to persist after the language has undergone sound change. They may tell us about the history of the language but not about how the language itself is represented in any speaker's brain.

It is possible, of course, to seek out or commission contemporary poetry from individuals who are unaware of or who have an imperfect knowledge of the traditional poetic conventions. Their use of rhymes, metrics, etc., should be a faithful reflection of their mental representation of words. As far as I know, this option has not been exploited by phonologists.

I therefore rate the evidence from poetry as 4·0, higher than the previously considered types of evidence – due to its potential, not its demonstrated usefulness to phonology.

2.4 Speech errors

Fromkin (1971) cites the speech error 'weeks and months' > '[wĩŋks] and . . .' as evidence for the psychological reality of the phonological rule which makes underlying nasals homorganic to following stops. This is a reasonable interpretation of such data if one can be sure that the nature of the error was simply the movement of the nasal from *months* into a similar environment (V—C) in *weeks*, where it then appears as a velar. But isolated speech errors may be subject to many alternative interpretations. *Winks* may have been the substitution of a whole lexical item (*winks* happens not only to be like *weeks* plus an inserted nasal, it is also an existing word). One may also ask whether the nasal consonant was actually

present. Malécot (1960) has shown that American English listeners (including, one assumes, those who collect speech errors) will accept a pronunciation of the sort [wĩks] for *winks*, i.e. with a nasalised vowel followed by an oral stop, even if the nasal consonant is completely absent. In this latter case, it would only be the feature of nasalisation which was taken from the word *months* and added to *weeks*. The point is, that given only the isolated speech error, typically gathered in completely natural settings (without being tape recorded) there is usually no way of verifying what the details of pronunciation were nor of deciding between competing analyses. Given a large number of speech errors, however, it may be possible to demonstrate statistically the plausibility of one way of analysing them over others since in this case the example-specific extraneous factors may cancel out. MacKay (1972), Shattuck-Hufnagel & Klatt (1979) and Shattuck-Hufnagel (this volume), among others, have made good use of this latter approach.

There is, to be sure, an inexhaustible supply of speech errors, although it may take a long time and considerable effort to gather a sufficient quantity of them or to find just the right type of error relevant to a given phonological issue. Thus, Fromkin (1971) originally claimed that phonotactic constraints were never violated in speech errors. New data show that this is not true, though (Shattuck-Hufnagel 1983: 130).

Speech errors do have the overriding advantage that they are completely natural and spontaneous and presumably give insight into the psychological structures and processes actually used by native speakers in the generation of speech. (This cannot be said of the structures and rules which allegedly underlie a speaker's knowledge of the relationship between *reside* and *residual*.)

For speech error data obtained under naturalistic conditions, then, I give the rating of 7·0. (Davidsen-Nielsen 1975; Shattuck-Hufnagel 1983; Baars *et al.* 1975; Motley & Baars 1975 have devised procedures which allow speech errors to be produced in quantity under controlled conditions. Stemberger & Lewis this volume make use of some of these methods. These techniques properly belong in the category of experiments which will be treated below.)

2.5 Word games

Word games which break words up into two or more parts and, optionally, rearrange them, permit the testing of a variety of phonological claims. Chao (1934), for example, demonstrated the allophonic variation in Chinese whereby /k/→[tɕ]/—/i y/ by showing that the word /mi/, when modified by a word game that inserts /aik/ after the initial consonant, becomes /meitɕi/ (the vowel change was not explained). Sherzer (1970) justified the analysis of the Cuna word [bíriga] 'year' (an exception to the otherwise fairly regular pattern of penultimate accent placement) as underlying /bírga/, since a word game which places the first syllable at the end transforms this word into [gabir] not *[rigabi], thus showing that

isation is a good candidate for psychologically motivated change and, furthermore, this has some empirical support (Bybee & Slobin 1982).

I therefore rate the evidence from sound change 2·0, a low score due to the extreme difficulty of making controlled observations of it.

2.3 Poetry

A song made popular by the singing group Peter, Paul, and Mary (and later John Denver) contains the lines in (5). Here *plane* [pʰlejn] is made to rhyme with *again* [əˈgɛn]:

(5) I'm leaving on a jet plane,
don't know when I'll be back again.

Phonetically the vowels in these two words obviously do not constitute a good rhyme (as pronounced by the singers). Could this, however, be used as evidence for these vowels being identical in their underlying (psychological) form? It might be if we could be sure of the source of knowledge the composer drew on when writing these lines. If established English-language conventions of rhyming prompted this choice of words, then the data contribute no insight into their underlying forms. (Regarding the quality of the second vowel in *again*, see Mencken 1948: 75–76.) Poetic conventions regarding permissible rhymes, metrics, etc., are usually quite conservative and once established tend to persist after the language has undergone sound change. They may tell us about the history of the language but not about how the language itself is represented in any speaker's brain.

It is possible, of course, to seek out or commission contemporary poetry from individuals who are unaware of or who have an imperfect knowledge of the traditional poetic conventions. Their use of rhymes, metrics, etc., should be a faithful reflection of their mental representation of words. As far as I know, this option has not been exploited by phonologists.

I therefore rate the evidence from poetry as 4·0, higher than the previously considered types of evidence – due to its potential, not its demonstrated usefulness to phonology.

2.4 Speech errors

Fromkin (1971) cites the speech error 'weeks and months' > '[wĩŋks] and . . . ' as evidence for the psychological reality of the phonological rule which makes underlying nasals homorganic to following stops. This is a reasonable interpretation of such data if one can be sure that the nature of the error was simply the movement of the nasal from *months* into a similar environment (V—C) in *weeks*, where it then appears as a velar. But isolated speech errors may be subject to many alternative interpretations. *Winks* may have been the substitution of a whole lexical item (*winks* happens not only to be like *weeks* plus an inserted nasal, it is also an existing word). One may also ask whether the nasal consonant was actually

present. Malécot (1960) has shown that American English listeners (including, one assumes, those who collect speech errors) will accept a pronunciation of the sort [wĩks] for *winks*, i.e. with a nasalised vowel followed by an oral stop, even if the nasal consonant is completely absent. In this latter case, it would only be the feature of nasalisation which was taken from the word *months* and added to *weeks*. The point is, that given only the isolated speech error, typically gathered in completely natural settings (without being tape recorded) there is usually no way of verifying what the details of pronunciation were nor of deciding between competing analyses. Given a large number of speech errors, however, it may be possible to demonstrate statistically the plausibility of one way of analysing them over others since in this case the example-specific extraneous factors may cancel out. MacKay (1972), Shattuck-Hufnagel & Klatt (1979) and Shattuck-Hufnagel (this volume), among others, have made good use of this latter approach.

There is, to be sure, an inexhaustible supply of speech errors, although it may take a long time and considerable effort to gather a sufficient quantity of them or to find just the right type of error relevant to a given phonological issue. Thus, Fromkin (1971) originally claimed that phonotactic constraints were never violated in speech errors. New data show that this is not true, though (Shattuck-Hufnagel 1983: 130).

Speech errors do have the overriding advantage that they are completely natural and spontaneous and presumably give insight into the psychological structures and processes actually used by native speakers in the generation of speech. (This cannot be said of the structures and rules which allegedly underlie a speaker's knowledge of the relationship between *reside* and *residual*.)

For speech error data obtained under naturalistic conditions, then, I give the rating of 7·0. (Davidsen-Nielsen 1975; Shattuck-Hufnagel 1983; Baars *et al.* 1975; Motley & Baars 1975 have devised procedures which allow speech errors to be produced in quantity under controlled conditions. Stemberger & Lewis this volume make use of some of these methods. These techniques properly belong in the category of experiments which will be treated below.)

2.5 Word games

Word games which break words up into two or more parts and, optionally, rearrange them, permit the testing of a variety of phonological claims. Chao (1934), for example, demonstrated the allophonic variation in Chinese whereby /k/→[tɕ]/—/i y/ by showing that the word /mi/, when modified by a word game that inserts /aik/ after the initial consonant, becomes /meitɕi/ (the vowel change was not explained). Sherzer (1970) justified the analysis of the Cuna word [bíriga] 'year' (an exception to the otherwise fairly regular pattern of penultimate accent placement) as underlying /bírga/, since a word game which places the first syllable at the end transforms this word into [gabir] not *[rigabi], thus showing that

the medial vowel is epenthetic. Hombert (1986) used word games in Bakwiri to demonstrate the non-segmental character of tone and vowel length in that they could be dissociated from the segments they are normally superimposed on. For example, /lùùŋgá/ 'stomach', when the syllables are interchanged, becomes /ŋgààlú/.

Evidence from word games has many of the same problems as evidence from poetry: one cannot always be sure whether the patterns they generate are purely conventional, i.e. culturally dictated behaviour, or whether they do indeed spring from the speaker's psychological representation and processing of words. And even if they do derive from psychological constructs, it is not always clear what 'level' of analysis the speakers operate on: surface or underlying. Carefully used, however, word games can aid in differentiating between competing hypotheses and it is a source of evidence that readily permits further refinements. It is a simple matter to elicit more of the same type of data.

For these reasons, I rate the evidence from word games at 8·0. (Some investigators have made attempts to teach speakers new word games in order to study specific phonological questions: M. Ohala 1975; Hombert 1986; Campbell 1986. These are essentially experiments and fall under the next category of evidence.)

2.6 Experiments

There is a popular misconception that what characterises experiments is the use of instruments, complex statistics, and the like. But, as with other evidence-gathering activities discussed above, the basic activity in experiments is *observing*. What differentiates the observations made during an experiment from those in the other situations is that in the former some care is taken to make them under circumstances which eliminate all anticipated potential distortions that might render their evidential value ambiguous. Typically this is done by directly *contriving the situation under which the observation is made*. In this way experiments can reduce the ambiguity of the evidence to a minimum. If someone can show that the experimental results fail to decide between the winner of the last contest and a new hypothesis not previously considered, a new experiment can be run. There is, therefore, no possibility of exhausting the evidence obtainable through experiments.

A common objection to experiments is that the contrived character of the circumstances under which the observations are made render their evidence less valid than that gathered in more natural settings. But if the feature of the experiment that is suspected of distorting the data is identified, then a new experiment can be designed that controls for this source of error. If the presence of a tape recorder is claimed to make subjects self-conscious and thus give unnatural responses to the experimenter's questions, then a hidden microphone can be used. If the laboratory environment is to blame, then the studies can be done in the street or in subjects' homes. And so on. Given sufficient imagination and resources, it is possible to artificially contrive any desired degree of 'naturalness' for the experiment; instructive examples may be found in the work of Pickett & Pollack (1963) and Kraut & Johnston (1979).

In contrast, when the experimental design is left up to nature, as in the case of poetry and the like, nature does not typically bother with proper experimental controls and as a result the evidential value of the data is less certain.

A casual inspection of the phonological literature shows that the experimental method has not been enthusiastically embraced by phonologists. This is rather puzzling, since the vast majority of phonologists are academics who regularly use the equivalent of experiments as part of the practice of teaching. Among their other duties, teachers are constantly required to certify that students know, i.e. are competent in, certain subject matters. Teachers do not assume that just because students have been exposed to the subject matter they therefore have mastered it. In other words, unlike general phonological practice, they do not assume that learners can automatically extract maximal generalisation from the data they are exposed to. Rather, teachers contrive various situations in which students can demonstrate their knowledge in an unambiguous way, e.g. performance on tests, production of an original piece of research, etc. Tests in the classroom, like tests in science, are never perfect, of course, and take considerable effort and imagination to construct, but conscientious teachers constantly strive to improve their tests. A practice that has worked well in academia should work in phonology.

Experiments, therefore, deserve the highest rating 9·5. (A perfect score of 10 would imply evidence that yields perfect knowledge, but this is unattainable by empirical means; see Ohala in press; Ohala & Jaeger 1986a).

2.7 Conclusions based on subjective ratings of evidence sources

The above 'consumer ratings' of sources of evidence used in phonology were given to be provocative – to encourage phonologists to evaluate soberly how well the evidence they are asked to swallow (or expect others to swallow) really helps to differentiate between all imaginable competing hypotheses. If pressed, I would be unable to justify all details of these ratings. The superiority of experimental evidence over the other types is still justified, though, if not by my arguments, then by the testimony of the history of science. Physics, chemistry and biology first became mature disciplines (with an accompanying marked increase in the rate of successful applications of their theories) when they started relying on and insisting on experimental evidence for claims. Physics may have been the first field to adopt this practice (in the sixteenth and seventeenth centuries) because the phenomena it first studied were relatively simple and permitted the implementation of controls more easily. Chemistry and biology, working with more complex systems, didn't see the benefit of experimentation until the nineteenth century. Human behaviour, including linguistic behaviour,

is certainly one of the most complex subject matters ever to be tackled by science, since it is caused by a host of factors which are not always easy to control when only one of them is the focus of inquiry. But the complexity of the problem does not by itself exempt us from using experiments in our investigation because as Claude Bernard ([1865] 1957: 2–3) remarked, addressing the same misconception in physiology:

> Experimentation is undeniably harder in medicine than in any other science; but for that very reason, it was never so necessary, and indeed so indispensable. The more complex the science, the more essential is it, in fact, to establish a good experimental standard, so as to secure comparable [i.e. reliable] facts, free from sources of error.

In other words, the more complex the science, the more possibility there is of dreaming up incorrect hypotheses and the most effective way of weeding these out is by resorting to experiments. This principle also suggests the reply to objections of the sort: 'Experimental methods lag behind developments in phonological theory so much as to render them useless to most practising phonologists. We don't have any way of testing such claims as the "strict cycle condition", that there is a hierarchical metrical structure underlying utterances, etc. If we had to wait for experimental support for these notions, we would never make any progress.' In fact, the evidence that claims of this sort are based on is equivocal – we have only to see the scores of theories spun out of the same linguistic data to be convinced of this – and so the 'progress' made is probably an illusion. Real progress, as Bernard suggests, would follow if an equal amount of imagination and enthusiasm were to be spent in the design and conduct of experiments as is currently spent in formulating the hypotheses that require testing.

Nevertheless, in spite of the inspiring and persuasive writings of scientists like Bernard, it must be admitted that it was the success of the experiments themselves done by him, Helmholtz, Pasteur and others which made the most converts to the experimental method. For this reason, in the remainder of this paper and in the other papers of this part of the *Phonology Yearbook*, examples are given of phonological experiments which demonstrate their potential for providing crucial evidence which helps to differentiate between competing hypotheses.

3 Phonological experiments

3.1 The psychological representation of MSCs

Every language creates its morphemes out of variable-length strings consisting of permutations of a small number of phonemes. Nevertheless, in every language certain permutations are not utilised; these are the phonotactic constraints or morpheme structure conditions (MSCs). In early generative phonology (EGP), e.g. Halle (1959), it was claimed that

Candidate word...	[kræk]	[klæb]	[klɛb]	[ðgøx]
Type of substitution				
CCVC	crack	–	–	–
CCVC	track	slab	–	–
CCVC	clack	crab	–	–
CCVC	creek	club	club	–
CCVC	crass	clam	Clem	–
CCVC	smack	stab	–	–
CCVC	click	crib	crib	–
CCVC	cream	clip	clip	–
CCVC	shriek	slob	slob	–
CCVC	clash	slam	phlegm	–
CCVC	trash	cram	Kress	–
CCVC	slick	grub	grub	–
CCVC	clean	cream	cream	–
CCVC	trim	flip	flip	–
CCVC	blab	spat	bread	–
CCVC	gleam	gruff	gruff	gruff
Total successful substitutions	16	15	12	1

[Table I. *Greenberg & Jenkins' phoneme-substitution algorithm for measuring the distance of words from the native pattern*]

MSCs were part of the derivative knowledge of the native speaker and that their function was to capture the significant generalisations about morpheme well-formedness. Thus the fact that, in English initial consonant clusters, a stop may not be followed by another obstruent would be reflected in an MSC but the fact that if the initial cluster of a monosyllable is /kl/ and the final consonant is /b/, the vowel may not be /ɛ/, would *not* merit representation in an MSC. The latter constraint would be a 'particularisation' rather than a generalisation.

Greenberg & Jenkins (1964, henceforth G&J) suggested that native speakers' knowledge of the phonotactics of their language could be modelled by a phoneme substitution algorithm which is illustrated in Table I as it applies to four CCVC strings, /kræk/, /klæb/, /klɛb/, /ðgøx/. In essence, it suggests that the degree of adherence of a given phoneme string to that of the native pattern is inversely proportional to the number of 1-, 2-, and up to *n*-phoneme substitutions for the original *n* phonemes of the string which succeed in yielding an existing word (or morpheme) in the language. G&J's hypothesis differs from that of EGP in that it predicts that speakers have an awareness of the *degree* to which a potential morpheme adheres to the native pattern and that any deviation, no matter how specific or how general, would be detectable. As shown in the table, /klɛb/ would be farther from English than /klæb/, even though neither could be said to violate an MSC as EGP would conceive of such.

Ohala & Ohala (1986) conducted an experiment to find out which of these hypotheses made better predictions regarding native speakers' reactions to certain made-up words. The experiment was patterned after

is certainly one of the most complex subject matters ever to be tackled by science, since it is caused by a host of factors which are not always easy to control when only one of them is the focus of inquiry. But the complexity of the problem does not by itself exempt us from using experiments in our investigation because as Claude Bernard ([1865] 1957: 2–3) remarked, addressing the same misconception in physiology:

> Experimentation is undeniably harder in medicine than in any other science; but for that very reason, it was never so necessary, and indeed so indispensable. The more complex the science, the more essential is it, in fact, to establish a good experimental standard, so as to secure comparable [i.e. reliable] facts, free from sources of error.

In other words, the more complex the science, the more possibility there is of dreaming up incorrect hypotheses and the most effective way of weeding these out is by resorting to experiments. This principle also suggests the reply to objections of the sort: 'Experimental methods lag behind developments in phonological theory so much as to render them useless to most practising phonologists. We don't have any way of testing such claims as the "strict cycle condition", that there is a hierarchical metrical structure underlying utterances, etc. If we had to wait for experimental support for these notions, we would never make any progress.' In fact, the evidence that claims of this sort are based on is equivocal – we have only to see the scores of theories spun out of the same linguistic data to be convinced of this – and so the 'progress' made is probably an illusion. Real progress, as Bernard suggests, would follow if an equal amount of imagination and enthusiasm were to be spent in the design and conduct of experiments as is currently spent in formulating the hypotheses that require testing.

Nevertheless, in spite of the inspiring and persuasive writings of scientists like Bernard, it must be admitted that it was the success of the experiments themselves done by him, Helmholtz, Pasteur and others which made the most converts to the experimental method. For this reason, in the remainder of this paper and in the other papers of this part of the *Phonology Yearbook*, examples are given of phonological experiments which demonstrate their potential for providing crucial evidence which helps to differentiate between competing hypotheses.

3 Phonological experiments

3.1 The psychological representation of MSCs

Every language creates its morphemes out of variable-length strings consisting of permutations of a small number of phonemes. Nevertheless, in every language certain permutations are not utilised; these are the phonotactic constraints or morpheme structure conditions (MSCs). In early generative phonology (EGP), e.g. Halle (1959), it was claimed that

Candidate word...	[kræk]	[klæb]	[klɛb]	[ðgøx]
Type of substitution				
CCVC	crack	–	–	–
CCVC	track	slab	–	–
CCVC	clack	crab	–	–
CCVC	creek	club	club	–
CCVC	crass	clam	Clem	–
CCVC	smack	stab	–	–
CCVC	click	crib	crib	–
CCVC	cream	clip	clip	–
CCVC	shriek	slob	slob	–
CCVC	clash	slam	phlegm	–
CCVC	trash	cram	Kress	–
CCVC	slick	grub	grub	–
CCVC	clean	cream	cream	–
CCVC	trim	flip	flip	–
CCVC	blab	spat	bread	–
CCVC	gleam	gruff	gruff	gruff
Total successful substitutions	16	15	12	1

[Table I. *Greenberg & Jenkins' phoneme-substitution algorithm for measuring the distance of words from the native pattern*]

MSCs were part of the derivative knowledge of the native speaker and that their function was to capture the significant generalisations about morpheme well-formedness. Thus the fact that, in English initial consonant clusters, a stop may not be followed by another obstruent would be reflected in an MSC but the fact that if the initial cluster of a monosyllable is /kl/ and the final consonant is /b/, the vowel may not be /ɛ/, would *not* merit representation in an MSC. The latter constraint would be a 'particularisation' rather than a generalisation.

Greenberg & Jenkins (1964, henceforth G&J) suggested that native speakers' knowledge of the phonotactics of their language could be modelled by a phoneme substitution algorithm which is illustrated in Table I as it applies to four CCVC strings, /kræk/, /klæb/, /klɛb/, /ðgøx/. In essence, it suggests that the degree of adherence of a given phoneme string to that of the native pattern is inversely proportional to the number of 1-, 2-, and up to *n*-phoneme substitutions for the original *n* phonemes of the string which succeed in yielding an existing word (or morpheme) in the language. G&J's hypothesis differs from that of EGP in that it predicts that speakers have an awareness of the *degree* to which a potential morpheme adheres to the native pattern and that any deviation, no matter how specific or how general, would be detectable. As shown in the table, /klɛb/ would be farther from English than /klæb/, even though neither could be said to violate an MSC as EGP would conceive of such.

Ohala & Ohala (1986) conducted an experiment to find out which of these hypotheses made better predictions regarding native speakers' reactions to certain made-up words. The experiment was patterned after

Word pair A	B	EGP's prediction	G&J's prediction	No. A responses	No. B responses
klɛb	klæb	A = B	A < B	4	12
θlɛz	θlɛd	A = B	A < B	1	15
ʃɹiz	ʃɹɪd	A = B	A < B	5	11
flɔt	flʊg	A = B	A < B	7	9
sθɛʃ	sθip	A = B	A < B	7	9
			Total	24	56
			($P \ll 0{\cdot}001$)		

[Table II. *First MSC experiment*]

that done earlier by Zimmer (1969) in his test of the psychological reality of Turkish MSCs. Pairs of words such as /klæb/ and /klɛb/ were constructed which should be judged as having different 'distances' from the native pattern of English if G&J are right but which should be judged equally far from English if EGP is right. Five such pairs (see Table II) were presented in random order to 16 American English-speaking subjects (linguistically naive students and staff at the University of California, Berkeley) via tape recorder or earphones, along with other pairs not relevant to the point under discussion. Subjects were asked to indicate by marking an answer sheet which member of each pair sounded more 'English-like'.

The results are shown in the two rightmost columns of Table II. Subjects' choices conformed to the predictions of the G&J hypothesis in 70% of the cases, which is highly significant statistically.

Assuming these results can be replicated, we may tentatively conclude that on those points which were the focus of the test, the G&J model is a better approximation than the EGP model in characterising the way native speakers react to candidate morphemes. This doesn't prove the correctness of the G&J model, of course. On the face of it, the constraints in their model of allowing only substitutions for phonemes and not additions or subtractions, and of permitting only consonant substitutions for consonants and vowels for vowels, seem arbitrary.[1] Moreover, as G&J note, the scoring procedure would probably be improved if the featural content of the phonemes substituted were taken into account.

An important aspect of the G&J model is that it requires only the lexicon (plus a means of accessing its contents) and some very general data processing mechanisms (the latter of which might serve for comparison of images, events, and other non-linguistic phenomena). The data processor does not contain and need not contain any abstracted ('derivative') knowledge about language-specific or language-universal sound patterns. Rules or generalisations such as 'in word-initial obstruent clusters, the first consonant must be /s/' would have no psychological reality in their model, rather only the words in the lexicon that adhere to this constraint will be psychologically real. It is interesting to speculate how much of speakers' phonological knowledge – not just MSCs – may be based only on their knowledge of the lexicon plus the possession of very general cognitive abilities.

Word	Response	Frequency of occurrence
scepticism	[skɛptɪk + ɪzm]	11
	[skept + ɪk + ɪzm]	1
	[skeptɪk + sɪzm]	13
	[skɛptə + sɪzm]	
	[skɛptə + ɪzm]	
plasticise	[plæstɪk + aĭz]	7
	[plæstɪs + aĭz]	1
	[plæstɪk + saĭz]	17
	[plæstə + saĭz]	
	[plæstə + aĭz]	
	one unit	
medicate	[mɛdɪk + eĭt]	5
	[mɛdə + keĭt]	20
	[mɛdə + eĭt]	
	[mɛd + keĭt]	
	[mɛd + ɪkeĭt]	
	[mɛdəsɪn + keĭt]	
	one unit	

[Table III. '*Meaningful parts*' *of derived words*]

3.2 Morphemic constituents of phonetically complex derivations

It is generally assumed by most phonologists today that native speakers are aware of the morphemic constituents of complex derived words, even those where the presumed stem is phonetically different from its form in other words, e.g. that *extrusion* consists of *extrude* plus *-ion*, and the like. Are they, though?

To test this, in a preliminary way, I first trained 25 linguistically naive English speakers (Berkeley students) to identify the 'meaningful parts' of morphemically complex words presented orally. I used simple examples in the training, e.g. pointing out that *yellowish* consisted of two meaningful parts, *yellow* and *ish*, both of which contribute to the meaning of the whole word but that in *furnish* there is only one meaningful part – the *ish*, in this case, does not independently contribute to the meaning of the whole, and so on. After they had demonstrated that they had correctly grasped the notion of 'meaningful part' (= 'morpheme') and could correctly apply it to various words where the constituent morphemes would always have an invariant shape, they were asked to do the same analysis to words such as *scepticism*, *plasticity*, and *medicate*. The results for these three words are given in Table III.

A majority of the responses to these words were of the type that revealed no clear awareness of the morphemic constituents in a way that is consistent with their use of phonological rules claimed to mediate between these and related forms, e.g. *sceptic*, *plastic*, *medicine*. It might be objected that the subjects did not really understand the nature of the task, but this is countered by the fact that they performed almost flawlessly on derivations which involved no or few phonetic changes in the constituents (*vis-à-vis* their appearance in other words), e.g. *highness*, *insuperable*, *retroactive*. Granted, with more careful training (such as that which we give to

students in linguistics classes) it would be possible eventually to turn linguistically naive subjects into linguistically trained subjects and have them give judgements that matched those of professional linguists. It seems fair to conclude, though, that speakers have more trouble recognising the constituent morphemes in derivationally complex words such as *plasticise* than they do in derivationally simple ones such as *highness* (see also McCawley this volume; Ohala & Ohala 1986). This undermines the assumptions commonly made by various recent schools of phonology. It suggests that if speakers are aware of a relationship between words like *sceptic* and *scepticism* this awareness may not include the kind of phonological knowledge needed to derive one from the other.

The above result should not be surprising; there are a number of cases where comparable misanalyses have permanently influenced the English vocabulary. *Witticism*, a coinage by Dryden (according to the OED), is based on *witty*+*ism*; where did the stem-final /s/ come from? Obviously the word is formed on the analogy of *critic/criticism* but the suffix in the derived form was taken to be [sɪzm] and so that is what was added to *witty*. Likewise, the /l/ in *Congolese* is based on a misanalysis of the stem + suffix boundary in the model for the derivation, *Angolese* (Malkiel 1966; see also Marchand 1969: 391ff).

When the phonetic shape and especially the meaning of a morpheme diverges too greatly in different words, e.g. in *purge/purgatory*, *labour/laboratory*, there is probably little motivation to a language user to keep track of what the original (i.e. 'underlying') parts are and where their boundaries lie, since there is already a great deal of purely idiosyncratic information about the words as wholes that must be memorised in any case. This reinforces the speculation made above that it is probably a mistake (without strong evidence) to make extravagant assumptions about the depth and detail of native speakers' knowledge of sound patterns in their language.

3.3 Underlying or surface segments?

It is well recognised that some variation in pronunciation stems from the physical constraints of the speech production system and is not a direct product of mental structures or processes. For example, the small Fo perturbations on vowels following voiced and voiceless consonants are thought to be a mechanical consequence of the gestures proper to the voicing distinction and are not purposeful (Hombert *et al.* 1979).

Another example of this sort is the epenthetic stops that are occasionally heard in words such as *teamster* ['tʰĩmpstɚ] and *youngster* ['jʌ̃ŋkstɚ]. These occur due to anticipatory assimilation, during the production of the nasal consonant, of the velic closure required by the [s], i.e. due to anticipatory denasalisation of the nasal (Grandgent 1896). On the other hand, such variations may, through sound change, become fossilised, i.e. entered in the speaker's mental lexicon. This is what accounts for the ⟨p⟩ in the conventional spelling of words such as *Thompson* (< *Thom*+*son*) and

dempster (< *deem*+*ster*). Much difference of opinion exists on how to treat such epenthetic stops: some regard them as surface phonetic phenomena (i.e. consequences of the physical constraints of the vocal tract and thus not a rule of grammar) and others as the product of 'higher' cognitive processes (e.g. phonological rules which insert the stops) (see Harms 1973; Donegan & Stampe 1979).

If we can assume that the durations of certain speech segments are determined by the phonemic composition of words, then measurement of these durations may reveal underlying phonemic make-up. It is well documented that vowel + nasal sequences are longer in open syllables than in syllables closed by obstruents, especially voiceless stops (Lovins 1978). So, if the duration of the [ĩm] sequence in a word such as *teamster* (with an epenthetic [p]) is unusually short it would imply that the [p] is the product of higher-level constructs (i.e. present at a stage of production planning prior to the processes that determine VN duration), otherwise not.

To demonstrate that durational measurements can differentiate between higher-level intended [p]'s and lower-level phonetically caused [p]'s, I ran the following study. I got 25 English-speaking subjects to speak the words *clam* and *clamp* and to derive them with the suffix *ster* (among other new derivations, such that the fact that these two forms constituted a minimal pair was presumably obscured). I assumed that any [p] found in the derivation *clam*+*ster* had to be a 'surface phonetic event' or at least couldn't have been lexicalised yet through sound change (as would be the case with the [p] in *Thompson*). The durations of the VN sequence in the *ster*-derived forms is given in Table IV. The length of the VN sequence was virtually the same in *clam*+*ster* whether it had a detectable epenthetic [p] or not; these durations, however, as predicted, were significantly greater than those in *clamp*+*ster*, where the [p] was intentional.

I next attempted to determine whether a [p] in a word like *teamster* was epenthetic or not in the pronunciation of a given speaker by examining the duration of that speaker's VN sequence in that word in comparison with that in *team* and in the forms *cam* and novel derived forms *cam*+*ster* and *camp*+*ster* (used instead of *clamp*, etc., to avoid potential ambiguity in segmentation between the [l] and the [æ̃]). I employed the following reasoning (where 'team', etc. stands for the 'duration of the VN portion of *team*'):

(6) If [p] is a surface event, then

$$\frac{\text{team}}{\text{team[p]ster}} = \frac{\text{cam}}{\text{cam}+\text{ster}}$$

whereas if [p] is intended, then

$$\frac{\text{team}}{\text{team[p]ster}} = \frac{\text{cam}}{\text{camp}+\text{ster}}$$

Word type	No. tokens	Mean	S.D.
clam[p]ster	8	350·7	54·0
clamster*	17	349·7	47·2
clamp + ster	24	255·1	33·2

* No clear epenthetic [p].

[Table IV. *Durations (in msec) of VN sequence in* clam + ster, clamp + ster]

cam/cam + ster	= 2·1
cam/camp + ster	= 2·49
team/team[p]ster	= 1·7

[Table V. *Duration ratio of VN sequences to test for epenthetic [p] (1 subject)*]

Typical results for *teamster* from one subject are given in Table V. For this speaker the ratio of the durations of the VN sequence were closer to those in *cam/cam + ster* than in *cam/camp + ster*, thus suggesting that the [p] was a purely surface event (see Ohala 1981b, c for further details).

3.4 The concept formation technique

One of the potentially most powerful experimental methods for the investigation of speakers' knowledge of the sound patterns in their language and at the same time one of the easiest to implement is that known as the 'concept formation' (CF) technique (see Derwing 1973; Baker *et al.* 1973; Jaeger 1986, this volume; Wang & Derwing this volume). Briefly, it involves 'teaching' subjects certain linguistic categories by providing them with examples of items from the various categories. No technical jargon or formal definitions are needed to convey to subjects a sense of the categorisation criteria.

Jaeger (1980) used the CF technique to answer a classic phonological question: do English speakers regard the voiceless unaspirated stops that appear after syllable-initial /s/ to be allophones of the voiced or voiceless stop phonemes, i.e. is the stop in *skill* categorised with the initial stop in *kill* or *gill* – or neither or both? The question has more than the usual interest surrounding the assignment of phones to phonemes since although most analysts assign this sound to the same phoneme category as the stop in *kill*, perception tests show that, stripped of its initial /s/, *skill* is perceived not as *kill* but as *gill*, apparently since the words of the latter type often begin with voiceless stops (Lotz *et al.* 1960). Jaeger wanted to see if linguistically naive subjects agreed with the traditional analysis or were more influenced by perceptual similarity of phones. In the training session of her experiment subjects listened to a list of words and were taught to differentiate those containing [k^h] (category) from those without it, including those having [g] (non-category). To steer subjects away from using an orthographic image of the words, the items identified as category

included words where the sound [k^h] was represented in conventional spelling with different letters, e.g. *kill*, *choir*, *quick*, *cash*. Further, words which would be spelled with some of these same letters appeared in the non-category exemplars, e.g. *knighthood*, *chip*, *cerebral*. In the test session, where no feedback was given, subjects overwhelmingly put words such as *skill* in the target category. Their categorisations therefore were in agreement with phonologists' traditional phonemic groupings of allophones.

It could be asked, however, whether there could be some small phonetic difference between [g̥] and [k] which influenced subjects to put them in different categories or whether subjects could have had some sort of response bias such that they automatically put the [k] or any new sound remotely like [k^h] into the target category and that they would do the same if the target category was /g/ (i.e. including both [g] and [g̥]). (I use the symbol [k] here only as the accepted phonetic notation for the voiceless velar stop, not to pre-judge the issue of its phonemic categorisation.) In an attempt to get answers to these questions, I conducted the following multi-part experiment.

Forty adult English-speakers (Berkeley students, as before), who were paid for their services, served as subjects and were assigned on the basis of their performance in a pre-test to one of four groups consisting of 10 subjects each. The assignment was done such that the average ability of subjects to perform on tests of this sort would be equal. All tests were conducted orally. The target category of the pre-test was 'noun'. In the main test Group 1's target category was words containing [k^h] or [sk]; Group 2's was words containing [g], [g̥] or [sk]. There was no test session for these two groups, since what was of interest was how many trials each would require to learn their respective categories. Would it be as easy for subjects to put the stop in [sk] in the /g/ phoneme as in the /k/ phoneme?

Group 3's target category was words with [k^h] and included in the non-category items were words like *skill* with their initial [s] spliced off, i.e. they sounded very much like *gill*. (Other words of this sort were *skate*, *scold* and *school*, which minus their [s]'s would sound like *gate*, *gold* and *ghoul*, respectively.) The test session for Group 3 contained words similar to those in the training session (but not the same tokens, of course) in addition to the 'unknown' words, the intact versions of *skill*, etc. In spite of the voiceless unaspirated stop such as that in (*s*)*kill* being given as an example of a non-category item, would subjects nevertheless include this sound in the same category as [k^h] in the test sessions? Group 4's target category was [g], [g̥] and words with initial [sk] such as *skill* from which the [s] had been removed. This group's test session contained words of the same type as were presented in the training session with the unknown words being the intact [sk] words, *skill*, etc. Would subjects who had been trained to group items like (*s*)*kill* with /g/ nevertheless reject it from that category when it was heard as intact *skill*?[2]

Details on the structure of the various parts of this test are given schematically in Table VI.

Test part	Sub-part	Group 1 (n = 10)	Group 2 (n = 10)	Group 3 (n = 10)	Group 4 (n = 10)
Pre-test	Training	Category: noun, e.g. *lizard, jury* Non-category: verb, etc., e.g. *vanquish, select*			
	Test	Same as above			
Main test	Training	Category: [k^h], [sk] e.g. *cash, skill* Non-category: [g], [g] and non-velars, e.g. *gill, fish, lip*	Category: [g], [g], [sk] e.g. *gill, skill, aghast* Non-category: [k^h], non-velars, e.g. *cash, fish, lip*	Category: [k^h] e.g. *cash* Non-category: [g], including [k] from spliced [sk] clusters, plus non-velars, e.g. *golf, (s)kill, fish*	Category: [g], including [k] from spliced [sk] clusters, e.g. *gash, (s)kill* Non-category: [k^h], non-velars, e.g. *cash, fish*
	Test	None	None	As above plus unknown [k] in intact [sk] clusters, e.g. *skill*	As above plus unknown [k] in intact [sk] clusters, e.g. *skill*

[Table VI. *Structure of the concept formation test*]

It was assumed that subjects had correctly learned the category assigned to them when they answered 15 trials correctly in a row with two or fewer errors, although in the main test subjects had to go through at least 30 trials whether they reached this criterion or not. This latter feature guaranteed that they would be exposed to a sufficient number and variety of control examples, e.g. those which showed orthography to be irrelevant.

This list of words (75 for each training session; 25 in test sessions) was taped along with the appropriate feedback and presented to subjects individually over earphones. Subjects gave oral responses which were noted manually by the experimenter who simultaneously monitored what the subjects heard and how they responded (see Jaeger 1986). The format of the tape and how the subjects would interact with it is indicated in (7):

(7)

	Time ⟶					
	warning signal	stimulus	subject's response	feedback		warning signal
Tape:	[tone] — (0·5 sec silence)	— 'catalogue' —	(4 sec silence)	— 'yes' —	(2 sec silence)	— [tone]
Subject:			'yes'			

In reporting the results three quantitative measures are given. 'Trials to criterion' indicates how many trials subjects required to learn the category (as defined above). The 'number reaching criterion' indicates how many subjects reached criterion before all 75 exemplar words had been presented. The 'categorisation' score, which was computed on the basis of the responses in the test sessions only, is defined slightly differently for words that were of the same type which appeared in the training session as opposed to the unknown words. For the former, this score equals the number of 'correct' responses (positive responses to category items and negative to non-category) minus the number of

'incorrect' responses (negative to category, positive to non-category), divided by the total number of such responses (excluding 'no response'). In the case of the unknown words, the score equals the number of positive responses minus the number of negative responses divided by the total number of responses to such words (excluding 'no response'). The computation of these scores is summarised in (8):

(8) Categorisation score for words of type in training session

$$= \frac{\left(\begin{matrix}\text{No. of positive to category} + \text{No. of negative to non-category}\end{matrix}\right) - \left(\begin{matrix}\text{No. of negative to category} + \text{No. of positive to non-category}\end{matrix}\right)}{\text{Total such responses (excluding 'no response')}}$$

Categorisation score for unknown words

$$= \frac{\left(\begin{matrix}\text{No. of positive} - \text{No. of negative}\end{matrix}\right)}{\text{Total such responses (excluding 'no response')}}$$

These scores can range between 1·0 and −1·0; the closer it is to 1·0, the more subjects included these items in the target category; the closer it is to −1·0, the more they excluded them.

The results are given in Table VII.

Test part	Measure	Group 1		Group 2	Group 3	Group 4
Pre-test	Trials to criterion	19·33		19·1	19·3	19·3
Main test	Trials to criterion	22·67	← * →	45·38	26·5	24·4
	No. reaching criterion	9		8	10	10
	Category score, trained words				0·975	0·96
	Category score, 'unknown'				0·80	−1·00

* Difference significant ($P < 0{\cdot}01$), Mann–Whitney sum of ranks (U = 15·5).

[Table VII. *Results of concept formation test*]

The similarity of the groups' mean trials to criterion on the pre-test was artificially contrived by assigning subjects to the various groups depending on their performance on the pre-test; thus, it was possible to achieve considerable equality in the four groups for ability at this sort of test.

Group 2, which had to learn to categorise the stop in [sk] with [g], took twice as long to reach criterion as Group 1, which had to put this stop in the same category with [k^h]. (These means are based only on those who did reach criterion.) This difference is statistically significant. We may conclude that the post-[s] [k] is felt to belong with [k^h] more than with [g] or [g], in accord with the traditional phonemic analysis. Group 3 also judged words with [sk] clusters as belonging in the same category as [k^h], even though the same type of words, minus the initial [s], were presented as non-category exemplars in the training session. Similarly, Group 4 excluded words with [sk] from the category that included [g] and [g], even though the same type of words, with the [s] spliced off, was presented to them in the training session as exemplars of that category. It seems clear,

again, that the traditional grouping of allophones made by linguists analysing English is in agreement with native speakers' intuitions. We may also conclude that this grouping is not motivated by fine phonetic detail but rather a combination of gross phonetic detail plus distributional information. The same unaspirated stop was treated differently depending on whether it appeared in absolute initial position or in post-[s] position.

4 Conclusion

In this paper I have argued that although phonologists may be able to extract useful information from all sources of evidence on what speakers know about the sound pattern of their language and how they represent this knowledge, some sources are more useful than others. Experimental evidence is most efficient in allowing us to weed out non-competitive hypotheses so that we can spend our time, efforts and resources on the hypotheses better in accord with the known facts. To recall the simile introduced earlier, experimentation is like harvesting apples by shaking the apple tree. Moreover, experiments offer, potentially, an endless supply of evidence, limited only by the researcher's time and imagination, and the requisite material resources. I have given examples of a variety of experimental methods which were put to use to gain evidence on various phonological issues. Other methods exist as well. The literature on experimental phonology – even if we focus only on those investigating psychological issues – is quite extensive (e.g. in addition to works already cited, Thumb & Marbe 1901; Esper 1925; Brown & Hildum 1956; Berko 1958; Ladefoged & Fromkin 1968; Anisfeld & Gordon 1968; Anisfeld 1969; Moskowitz 1973; LaRiviere *et al.* 1974; Schane *et al.* 1974; Jarvella & Snodgrass 1974; Steinberg & Krohn 1975; Baker & Smith 1976; MacKay 1976, 1978; Myerson 1976; Fox & Terbeek 1977; Cena 1978; Stanners *et al.* 1979; Prideaux *et al.* 1980; M. Ohala 1983; Ohala & Jaeger 1986b; to mention just a few).

Our discipline should continue to augment the arsenal of experimental methods available to grapple with the issues that confront it. But it is also true that currently available methods can already provide solutions to many current issues if phonologists would just use them.

NOTES

* I gratefully acknowledge the helpful comments and criticisms of members of the audience at the Department of Linguistics, University of Michigan, especially R. Rhodes and P. Benson, where this paper was first presented. In addition I thank M. Caisse, J. Jaeger and S. Pearson for help and advice in the conduct of the experiments and J. Jaeger, C. Gussenhoven, G. Nathan and M. Ohala for insightful critiques of an earlier version of this paper. Errors and infelicities which remain are on my head. Portions of the research reported here were funded by the Committee on Research, University of California.

[1] Pertz & Bever (1975), in an interesting experiment, show that monolingual English-speaking children and adolescents judge made-up words such as *ntiff*, *ndall*, to be 'easier, more likely, or more usual' (i.e. in linguistic terms, 'unmarked') than words such as *nkiff*, *mdall*. They argue that these judgements must be based on knowledge of universal marking constraints (favouring homorganic nasal + stop clusters over heterorganic ones), which knowledge must not be based on an examination of sound patterns present in the English lexicon – since forms of the above sort are all equally far from the English pattern as measured by the G&J metric. If, however, the G&J algorithm is incorrect in not allowing additions, such that e.g. *ndall* could be changed into an existing word by the prefixation of [sæ] yielding *sandal*, then their conclusion may be questioned.

[2] A post-test was also given to all four groups. These results once fully analysed will be reported in a later paper.

REFERENCES

Anisfeld, M. (1969). Psychological evidence for an intermediate stage in a morphological derivation. *Journal of Verbal Learning and Verbal Behavior* **8**. 191–195.

Anisfeld, M. & M. Gordon (1968). On the psychophonological structure of English inflectional rules. *Journal of Verbal Learning and Verbal Behavior* **7**. 973–979.

Baars, B. J., M. Motley & D. G. MacKay (1975). Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior* **14**. 382–391.

Baker, W. J., G. D. Prideaux & B. L. Derwing (1973). Grammatical properties of sentences as a basis for concept formation. *Journal of Psycholinguistic Research* **2**. 201–220.

Baker, R. G. & P. T. Smith (1976). A psycholinguistic study of English stress assignment rules. *Language and Speech* **19**. 9–27.

Berko, J. (1958). The child's learning of English morphology. *Word* **14**. 150–177.

Bernard, C. (1957). *An introduction to the study of experimental medicine*. New York: Dover. (First French edition: 1865.)

Brown, R. W. & D. C. Hildum (1956). Expectancy and the identification of syllables. *Lg* **32**. 411–419.

Bybee, J. L. & D. I. Slobin (1982). Rules and schemas in the development and use of the English past tense. *Lg* **58**. 265–289.

Campbell, L. (1986). Testing phonology in the field. In Ohala & Jaeger (1986b). 163–173.

Cena, R. M. (1978). When is a phonological generalization psychologically real? Indiana University Linguistics Club.

Chao, Y.-R. (1934). The non-uniqueness of phonemic solutions of phonetic systems. *Bulletin of the Institute of History and Philology* (Academia Sinica) **4**. 363–397.

Chomsky, N. & M. Halle (1968). *The sound pattern of English*. New York: Harper & Row.

Davidsen-Nielsen, N. (1975). A phonological analysis of English *sp*, *st*, *sk* with special reference to speech error evidence. *Journal of the International Phonetic Association* **5**. 3–25.

Derwing, B. L. (1973). *Transformational grammar as a theory of language acquisition*. Cambridge: Cambridge University Press.

Derwing, B. L. & W. J. Baker (1977). The psychological basis for morphological rules. In J. Macnamara (ed.) (1977). *Language learning and thought*. New York: Academic Press. 85–110.

Donegan, P. & D. Stampe (1979). The study of natural phonology. In D. A. Dinnsen (ed.) *Current approaches to phonological theory*. Bloomington: Indiana University Press. 126–173.

Esper, E. A. (1925). A technique for the experimental investigation of associative interference in artificial linguistic material. *Language Monograph* **1**.
Fox, R. A. & D. Terbeek (1977). Dental flaps, vowel duration and rule ordering in American English. *JPh* **5**. 27–34.
Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Lg* **47**. 27–52.
Grandgent, C. H. (1896). Warmpth. *Publication of the Modern Language Association* **11** (New Series **4**). 63–75.
Greenberg, J. H. & J. J. Jenkins (1964). Studies in the psychological correlates of the sound system of American English. *Word* **20**. 157–177.
Halle, M. (1959). *The sound pattern of Russian*. The Hague: Mouton.
Harms, R. T. (1973). Some non-rules of English. Indiana University Linguistics Club.
Hombert, J.-M. (1986). Word games: some implications for analysis of tone and other phonological processes. In Ohala & Jaeger (1986b). 175–186.
Hombert, J.-M., J. J. Ohala & W. G. Ewan (1979). Phonetic explanations for the development of tone. *Lg* **55**. 37–58.
Jaeger, J. J. (1980). Testing the psychological reality of phonemes. *Language and Speech* **23**. 233–253.
Jaeger, J. J. (1984). Assessing the psychological status of the Vowel Shift Rule. *Journal of Psycholinguistic Research* **13**. 13–36.
Jaeger, J. J. (1986). Concept formation as a tool for linguistic research. In Ohala & Jaeger (1986b). 211–237.
Jarvella, R. J. & J. G. Snodgrass (1974). Seeing *ring* in *rang* and *retain* in *retention*: on recognizing stem morphemes in printed words. *Journal of Verbal Learning and Verbal Behavior* **13**. 590–598.
Kahn, D. (1976). *Syllable-based generalizations in English phonology*. Indiana University Linguistics Club.
Kiparsky, P. (1968). Linguistic universals and linguistic change. In E. Bach & R. T. Harms (eds.) *Universals in linguistic theory*. New York: Holt, Rinehart & Winston. 170–202.
Kraut, R. E. & R. E. Johnston (1979). Social and emotional messages of smiling: an ethological approach. *Journal of Personality and Social Psychology* **37**. 1539–1553.
Ladefoged, P. & V. A. Fromkin (1968). Experiments on competence and performance. *IEEE Transactions on Audio and Electroacoustics* **AU-16**. 130–136.
LaRiviere, C., H. Winitz, J. Reeds & E. Herriman (1974). The conceptual reality of selected distinctive features. *Journal of Speech and Hearing Research* **17**. 122–133.
Lotz, J., A. S. Abramson, L. J. Gerstman, F. Ingemann & W. J. Nemser (1960). The perception of English stops by speakers of English, Spanish, Hungarian, and Thai: a tape-cutting experiment. *Language and Speech* **3**. 71–77.
Lovins, J. B. (1978). 'Nasal reduction' in English syllable codas. *CLS* **14**. 241–253.
MacKay, D. G. (1972). The structure of words and syllables: evidence from errors in speech. *Cognitive Psychology* **3**. 210–227.
MacKay, D. G. (1976). On the retrieval and lexical structure of verbs. *Journal of Verbal Learning and Verbal Behavior* **15**. 169–182.
MacKay, D. G. (1978). Derivational rules and the internal lexicon. *Journal of Verbal Learning and Verbal Behavior* **17**. 61–71.
MacNeilage, P. F. (ed.) (1983). *The production of speech*. New York: Springer-Verlag.
Malécot, A. (1960). Vowel nasality as a distinctive feature in American English. *Lg* **36**. 222–229.
Malkiel, Y. (1966). Genetic analysis of word formation. In T. A. Sebeok (ed.) *Current Trends in Linguistics* **3**. The Hague: Mouton. 305–364.
Marchand, H. (1969). *The categories and types of present-day English word-formation*. 2nd edn. Munich: Beck.

Mencken, H. L. (1948). *The American language*. Supp. 2. New York: Knopf.
Moskowitz, B. A. (1973). On the status of vowel shift in English. In T. E. Moore (ed.) *Cognitive development and the acquisition of language*. New York: Academic Press. 223–260.
Motley, M. T. & B. J. Baars (1975). Encoding sensitivities to phonological markedness and transition probability: evidence from spoonerisms. *Human Communication Research* **2**. 351–361.
Myerson, R. F. (1976). Children's knowledge of selected aspects of 'Sound pattern of English'. In R. N. Campbell & P. T. Smith (eds.) *Recent advances in the psychology of language: formal and experimental approaches*. New York: Plenum Press. 377–402.
Ohala, J. J. (1974). Experimental historical phonology. In J. M. Anderson & C. Jones (eds.) *Historical linguistics*. Vol. 2. Amsterdam: North-Holland. 353–389.
Ohala, J. J. (1981a). The listener as a source of sound change. In C. S. Masek, R. A. Hendrick & M. F. Miller (eds.) *Papers from the parasession on language and behavior*. Chicago: Chicago Linguistic Society. 178–203.
Ohala, J. J. (1981b). Speech timing as a tool in phonology. *Phonetica* **38**. 204–212.
Ohala, J. J. (1981c). Articulatory constraints on the cognitive representation of speech. In T. Myers, J. Laver & J. Anderson (eds.) *The cognitive representation of speech*. Amsterdam: North-Holland. 111–122.
Ohala, J. J. (1983a). The phonological end justifies any means. In S. Hattori & K. Inoue (eds.) *Proceedings of the 13th International Congress of Linguists, Tokyo*. Tokyo: distributed by Sanseido Shoten. 232–243.
Ohala, J. J. (1983b). The origin of sound patterns in vocal tract constraints. In MacNeilage (1983). 189–216.
Ohala, J. J. (in press). Explanation in phonology: opinions and examples. In W. U. Dressler (ed.) *Phonologica 1984*. Cambridge: Cambridge University Press.
Ohala, J. J. & J. J. Jaeger (1986a). Introduction. In Ohala & Jaeger (1986b). 1–12.
Ohala, J. J. & J. J. Jaeger (eds.) (1986b). *Experimental phonology*. Orlando, Fl.: Academic Press.
Ohala, J. J. & M. Ohala (1986). Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In Ohala & Jaeger (1986b). 239–252.
Ohala, M. (1975). Nasals and nasalization in Hindi. In C. A. Ferguson, L. M. Hyman & J. J. Ohala (eds.) *Nasálfest: papers from a symposium on nasals and nasalization*. Stanford: Language Universals Project. 317–332.
Ohala, M. (1983). *Aspects of Hindi phonology*. Delhi: Motilal Banarsidass.
Pertz, D. L. & T. G. Bever (1975). Sensitivity to phonological universals in children and adolescents. *Lg* **51**. 149–162.
Pickett, J. M. & I. Pollack (1963). Intelligibility of excerpts from fluent speech: effects of rate of utterance and duration of excerpt. *Language and Speech* **6**. 151–164.
Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.
Prideaux, G. D., B. L. Derwing & W. J. Baker (eds.) (1980). *Experimental linguistics: integration of theories and applications*. Ghent: E. Story-Scientia.
Saussure, F. de (1916). *Cours de linguistique générale*. Paris: Payot. Translated (1966) as *Course in general linguistics*. New York: McGraw-Hill.
Schane, S., B. Tranel & H. Lane (1974). On the psychological reality of a natural rule of syllable structure. *Cognition* **3**. 351–358.
Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In MacNeilage (1983). 109–136.
Shattuck-Hufnagel, S. & D. H. Klatt (1979). The limited use of distinctive features and markedness in speech production: evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior* **18**. 41–55.
Sherzer, J. (1970). Talking backwards in Cuna: the sociological reality of phonological descriptions. *Southwestern Journal of Anthropology* **26**. 343–353.

Stanners, R. F., J. J. Neiser, W. P. Hernon & R. Hall (1979). Memory representation for morphologically related words. *Journal of Verbal Learning and Verbal Behavior* **18**. 399–412.

Steinberg, D. D. & R. K. Krohn (1975). The psychological validity of Chomsky and Halle's Vowel Shift Rule. In E. F. K. Koerner (ed.) *The transformational-generative paradigm and modern linguistic theory*. Amsterdam: John Benjamins. 233–259.

Sweet, H. (1874). *History of English sounds*. London: Trübner.

Thumb, A. & K. Marbe (1901). *Experimentelle Untersuchungen über die psychologischen Grundlagen der sprachlichen Analogiebildung*. Leipzig: Wilhelm Engelmann.

Wang, H. S. (1985). *On the productivity of vowel shift alternations in English: an experimental study*. PhD dissertation, University of Alberta.

Zimmer, K. (1969). Psychological correlates of some Turkish morpheme structure conditions. *Lg* **45**. 309–321.