24. Experimental Phonology

JOHN J. OHALA

Subject	Theoretical Linguistics » Pholonogy
DOI:	10.1111/b.9780631201267.1996.00026.x

0 Introduction: The Basis of Experimental Phonology

Experimentation in phonology and indeed in any scientific discipline has two basic elements for its philosophical foundation: the first is doubt. Doubt or scepticism that perhaps the things we believe or what others would have us believe may not correspond to the way things are. We realize that perhaps the sources of our beliefs are imperfect: the authority or impartiality of our teachers (including those who teach us through the printed word) may be flawed and our own senses may give us distorted or contaminated information about the world. Such scepticism, of course, has been the basis of many religious and philosophical systems. In these other domains the typical response to suspect beliefs is the promotion of additional, different beliefs (which may themselves be suspect!). What experimental methods propose in response to doubt – and this is the second element – is that it is possible to do something to counteract the suspected sources of error.[1] Specifically, it should be possible to anticipate them and to eliminate them or at least to limit their influence. In sum, experimentation rests on the following equation: if one believes, one may doubt; if one doubts, one can strive to resolve the doubt. Part of the lore of every scientific discipline are the possible sources of error in interpreting uncontrolled observations and the procedures, often quite ingenious, for compensating for them.

Thus, it must be emphasized that an experiment is prompted by a belief, i.e., a hypothesis or theory, which is subject to reasonable doubt. There can be no true experiments without theories; they have to be done with a purpose. And conversely, the only beliefs not subject to doubt and thus never subjected to test are religious dogmas (but see Kings 3.18, 21–40 [Douay version]). Second, an experiment consists physically of an observation under contrived or controlled circumstances, the control or contrivance being such that would eliminate or attenuate some anticipated or suspected distortion in the prior observations or events that gave rise to the belief.

Claude Bernard (1957) differentiated between nature-made experiments and man-made experiments. In nature-made experiments, nature manipulates the variables and the only contrivance on the part of man is to be in the right place at the right time to make the relevant observations. In man-made experiments, the experimenter controls the variables and makes the observations. Man-made experiments are far more efficient ways to test theories since nature often does not oblige us by presenting situations where the variables of interest – and no other – are systematically manipulated. But man-made experiments often require considerable ingenuity so that the artificiality of the conditions of observation do not themselves introduce unacceptable distortions.

There are opportunities in phonology to take advantage of nature-made experiments: speakers making novel derivations (adding the suffix *-ity* to *mundane*), speech errors, naturally-occurring defects of the speech production and perception system (e.g., aglossia or the absence of a tongue). In this chapter, however, I consider only man-made experiments because they are, potentially at least, applicable to any theoretical issue – limited only by resources and the experimenter's ingenuity.

1 Experimentation in Phonology

No claims in phonology are above doubt: the existence of the phoneme, syllable, or the feature [voice]; the reconstructed Proto-Indo-European form for Sanskrit *budh-*; that speakers know the posited rule-governed phonological link between the pair of words *repose/repository*. All of these are potential subjects for experimental study. But experiments are expensive: in time, effort, and other resources. It is a matter of research strategy, the availability of reliable experimental methods, and the amount of personal commitment we have to one belief or another which determines which issues one chooses to address experimentally. It is also worth nothing that in every discipline, even those where a tradition of experimentation is well established, like chemistry and physiology, scientists engage in nonexperimental activities such as description and classification of observations, delving into the history of ideas and methods in their discipline, offering speculations that range from the "wild and wooly" to those bolstered by extensive and rigorous arguments. No discipline "closes up shop" just because experiments are not applied to every issue. It takes time to develop an arsenal of reliable experimental methods, and even after that is achieved they may not be applied because of expense or lack of interest. Nevertheless, adequate testing of claims remains a prerequisite to understanding.

So many experimental paradigms have been proposed for testing phonological hypotheses (Ohala and Jaeger 1986; Ohala 1986; Ohala and Ohala 1987; Prideaux, Derwing, and Baker 1980; Kingston and Beckman 1990; Docherty and Ladd 1992; Diehl 1992) that it is impossible to review them all in a single chapter. I can only give representative examples from different domains in phonology.

1.1 Experimental Assessment of the Distinctive Features of Speech

One of the most fundamental tasks of phonology is to establish how different linguistic messages are conveyed by sound. Whether it is lexical differences or grammatical function, distinct messages must have distinct physical encodings, whether these are paradigmatic (different ciphers from a finite inventory of ciphers) and/or syntagmatic (different permutations of the ciphers). This is far from a trivial issue and certainly not one to be determined unequivocally by the unaided ear. Well-established methods exist for discovering the physical correlates of different linguistic messages in cases where they are uncertain or disputed. Although such studies are often regarded as having purely phonetic, not phonological, interest, this is a mistake: without having an "anchor" in the real world, phonology risks having its claims apply only in an imaginary universe of little interest to those outside the narrowly circumscribed world of autonomous phonology. Fortunately, such a parochial view of phonology is disappearing.

Consistent differences may be sought in the physiological or acoustic domains but the relevance of any difference found must ultimately be validated in the perceptual domain (Lehiste 1970). For example, in a series of instrumental and experimental studies, Lisker and Abramson (1964, 1967, 1970) found that in initial position (before stress), the distinction between pairs of English words like *paid* vs. *bade, tie* vs. *die, cool* vs. *ghoul*, is carried largely by the relative timing of voice onset after the stop release, that is, what is called *VOT* (for *V*oice *O*nset *T*ime): the phonemes / p t k/ showed a substantial delay in VOT (modal VOT =50-70 msec) whereas / b d g / had a short VOT (modal VOT = 0-20 msec). Phonetically, this contrast is said to be between voiceless aspirated stops and voiceless unaspirated stops. Perceptual studies demonstrated that VOT was the dominant cue for such lexical distinctions although several secondary cues also played a role (Lisker 1986). Although this contrast among stops is commonly attributed to presence vs. absence of voice, voicing *per se* plays only a secondary role in this environment and in other positions in the word as well (Denes 1955; Raphael 1972). Lisker (1957) showed that in intervocalic position, in addition to voicing, the duration of stop closures helps to cue lexical distinctions such as *rapid* vs. *rabid*, where the voiceless stop is longer.

The stops that appear in prevocalic clusters after / s /, e.g., *spade, sty, school*, may only be voiceless unaspirated. Lotz et al. (1960) showed that to English speakers these are perceptually most similar to the stops in *bade, die*, and *ghoul*, i.e., /b d g/ (though they are not completely identical, Caisse 1981). Thus, although traditionally the prevocalic stops in *paid* and *spade* would be counted as allophones of the same phoneme / p / in English, there is greater physical and perceptual similarity between the stops in *bade* and *spade*.

1.2 Can Phonetically Different Sounds Be Psychologically the Same?

This still leaves open the question of whether native speakers regard the voiceless unaspirated stops in *s*C- clusters to be psychologically similar to the voiceless aspirated or the voiceless unaspirated in absolute initial position. This question was investigated by Jaeger (1980, 1986) who used the so-called concept formation method to address the question of the assignment of allophones to the /k/ phoneme. Without being given any more instructions than (approximately) "assign the following words to two different categories depending on the pronunciation at their beginning," linguistically naïve subjects were first presented orally with uncontroversial examples with initial stops such as *kiss, chasm, cattle*, and *quake* designated "category," intermixed randomly with noncategory examples, *grip, gash, lime, ceiling, chest,* and *knife*. Initially subjects were given feedback on each trial, i.e., they were told whether their category assignment was correct or not. If they reached some preset criterion of performance in this training, they were then presented with words containing the stop allophone whose phoneme membership was controversial, such as *school* and *scold*. This time there was no feedback. If they put these words in the same category as *cool* and *cold* it would imply that they regarded the [k] and [k^h] as somehow psychologically equivalent. In fact, this is what they did. (See also Ohala 1986.)

Jaeger attempted to control for bias from orthography (in case subjects visualized the spelling and let this influence their judgments): thus the words with [k^h] were spelled with varying letters, *k, c, ch, qu*, and some of the same letters started noncategory words. Nevertheless, it is difficult to control completely for othographic bias when using literate subjects. In fact, there is a growing body of evidence that much of what is regarded as native speakers' knowledge of the phonology of their language is very much influenced by, if not based on, their knowledge of how their language is spelled (Wang and Derwing 1986; Derwing and Nearey 1986; Read, Yun-fei, Hong-yun, and Bao-qing, 1986; Morais, Cary, Alegria and Bertelson 1979).

1.3 Experiments on Morpheme Struture Constraints

As it happens, some of the earliest linguistic and phonological experiments ever done were intended to address the issue of language change but in fact also gave evidence on the psychological processes underlying language use. Thumb and Marbe (1901) tested the posited effect of word association on language change. Inspired by this work, Esper (1925) explored the effect of analogy on the change in phonological shape of words and morphemes. His experiment was a task where he required his subjects to learn the names of 16 objects, each having one of four different shapes and one of four different colors. (He trained them on 14 object-name associations but tested them on 16 in order to see if they could generalize what they learned.) In three different experimental conditions, each with a different group of subjects, the relationship between the names and properties of the objects differed. The names presented to subjects in group 1 were of the sort *naslɔ*, *šownlɔ*, *nasdeg*, *šowndeg*, where *nas-* and *šown* coded color and *Lɔ* and *-deg* coded shape (though they were not told of their "morphemic" constituents). Since these names consisted of two phonologically legal morphemes, this group could simplify their task by learning not 16 names but 8 morphemes (if they could discover them) plus the simple rule that the color morpheme preceded the shape morpheme in each name. Group 3, a control group, were presented names that had no morphemic structure; they had no recourse but to learn 16 idiosyncratic names. As expected, group 1 learned their names much faster and more accurately than group 3. Of interest was the performance of group 2 which, like group 1, were presented with bi-morphemic names and thus could, in principle, simplify their task by learning just eight morphemes. But, unlike group 1, the morphemes were not phonologically legal for English, e.g., *nulgɛn, nuzgub, pelgɛn, pezgub* (where now *nu-* and *pe-* were color morphemes and -*lgɛn* and -*zgub* were shape morphemes, the latter two, of course, violating English morpheme structure constraints). Could the subjects in group 2 extract the hidden morphemes and perform as well as those in group 1? Apparently not: their performance was similar to (and marginally worse than) that of group 3, which had 16 idiosyncratic names to learn. Furthermore, analysis of the errors of group 2, including how they generalized what they'd learned to the two object-name associations excluded from the training session, revealed that they tried to make phonologically legal morphemes from the ill-formed ones. Esper's experiment achieved his goal of showing the force of analogy in language change, i.e., paradigm regularization, but it also demonstrates the psychological reality of morpheme structure constraints.

1.4 Experiments on Phonological Change

One of the earliest accomplishments of phonology was the development of a method, the comparative method, which allowed one to reconstruct the history of languages, in particular the changes over time in the phonological forms of words (Rask 1818; Grimm 1822). To oversimplify, the comparative method consists in finding an optimal single unbranching path between pairs or groups of words judged to be cognates, where the "path" consists of (a) intermediate forms between the two, one of which is the "parent form" and (b) sound changes which operate unidirectionally and convert the parent form into the attested daughter forms. Historical phonology might seem at first to be an unlikely domain for experimentation since most of the events of interest occurred in the inaccessible past and thus cannot be manipulated by the experimenter. But if one is willing to make the unformitarian assumption,[2] that is, that whatever caused sound changes in the past is still present and causing sound changes now, then although we cannot be there when Proto-Indo-European k^w changed to Greek p, e.g., PIE *ekwos* "horse" > Gk. *hippos*, we may be able to contrive circumstances where the same or similar changes occur in front of our eyes or our microphones. In fact, the parallelism between diachronic and synchronic variation has often been remarked by researchers and sometimes has led to laboratory-based studies of sound change (Rousselot 1891; Haden 1938). One of the most fruitful areas of experimental phonology, then, involves studies on the phonetic influence on sound change or on phonological universals in general (see, e.g., Lindblom 1984; Wright 1986; Kawasaki 1986; Kawasaki-Fukumori 1992; Stevens 1989; Goldstein 1983; Ohala 1992, 1993).

One of the most common processes evident in sound change is assimilation and one of the common textbook examples of it is the case of medial heterorganic clusters assimilating in Italian: Late Latin *octo* > Italian *otto* "eight". Such assimilations are overwhelmingly of the form $-C_1C_2- > -C_2C_2-$; rarely does C_2 assimilate to the place of C_1 (and many of these cases could be reanalyzed as involving a different process; see Murray 1982). Such a change is usually attributed to *ease of articulation* or conservation of energy (a heterorganic cluster requiring more energy than a homorganic one). But if so, why is it C_1 that usually changes, not C_2? Expenditure of articulatory energy is presumably cumulative through an utterance and thus would be greater by the time C_2 was reached than C_1. Thus we might expect C_2 to assimilate to C_1, just the reverse of what is found. Such doubts lead us to entertain an alternative explanation for this process. Ohala (1990) reported the results of an experiment designed to test whether the process might better be attributed to acoustic-auditory factors. This was an experiment where the two halves of VCV utterances ([apa, ata, aka, aba, ada, aga]) where separated at the middle of the stop closure and reattached via digital splicing in order to create a variety of VCV stimuli where the stop onset and stop release had different places, e.g., first part of [apa] spliced to last half of [ata] to yield [ap-ta], but the medial closure duration was that for a singleton stop. These were presented, randomized, to listeners who were asked to identify the stop, being allowed the options of reporting it as C_1, C_2, or "other." Of the tokens where $C_1 > C_2$, 93 percent of the responses were C_2. A subsequent test showed that if the stop closure interval were lengthened, eventually the majority of listeners could hear the heterorganic cluster but the threshold duration for this was longer for voiceless stops than voiced stops. Presumably listeners were influenced by their awareness that singleton voiceless stops are longer than voiced stops (see above). Plausibly the place cues for C_2 dominate over those for C_1 (even when they are inconsistent) because they are acoustically and auditorily more salient (and listeners learn where to invest most of their auditory attention): both onset and offset have some formant transitions which cue place but only the offset has the very important cues contained in the stop burst. Although there is still much to be learned about the historical processes that changed *octo* to *otto*, this study at least shows that there is plausibly a major acoustic-auditory component to it; appeals to "ease of articulation" may be unnecessary. It also demonstrates the potential for an experimental approach to questions in historical phonology.

1.5 Experiments in Lexical Representation

It was mentioned in note 1 above that tests may only be made of claims which involve things that ultimately, even if indirectly, have observable consequences. As soon as a claim is associated with observable consequences, it becomes testable. Lahiri and Marslen-Wilson (1991, 1992) put underspecification theory into the empirical arena. They suggested that the lexical representations posited by phonologists "correspond, in some significant way, to the listener's mental representation of lexical forms… and that these representations have direct consequences for the way … the listener interprets the incoming acoustic-phonetic information.” Lahiri (1991) argued specifically that “the surface structures derived after postlexical spreading do not play a distinctive role in perception; rather, a more abstract underspecified representation determines the interpretation of a phonetic cue.”

In English, vowels are not lexically specified for the feature [nasal], and thus an oral vowel heard without a following consonant is predicted to be ambiguous as to whether it is in a CVC or a CVN word.[3] Lahiri and Marslen-Wilson tested this using a paradigm that involves gating (truncating) the ends of words in increments of 40 msec and presenting the gated fragments to English-speaking subjects in the order of most to least gated, and then asking them to guess what the word is. Although 83.4 percent of the responses to the CVC stimuli to the point where the final consonant abutted the vowel were correct (i.e., CVC), Lahiri and Marslen-Wilson interpreted the 16.6 percent CVN-responses as consistent with the vowels being unspecified for [nasal] and thus ambiguous between their coming from CVC or CVN words. Figure 24.1 shows their subjects' responses as a function of the gating point (circles).[4] Lahiri and Marslen-Wilson did not perform any statistical analyses on their data and, of course, it would have been impossible given the open response set subjects could choose from; that is, to know whether the CVN responses occurred at a rate equal to, more than, or less than chance, one would have to know how many possible responses they could have made of each word type.

Ohala and Ohala (1993) attempted a replication of Lahiri and Marslen-Wilson's experiment,[5] but restricted the subjects' responses to just one of two choices, e.g., when presented with an end-gated version of *rube* the choices specified on the answer sheet were *room*. *rube*. The results are seen in figure 24.1 as triangles superimposed on the circles representing Lahiri and Marslen-Wilson. This curve appears to be quite similar to theirs, but there is a crucial difference: a statistical analysis is possible in the Ohala and Ohala case. In fact these results show that listeners made the correct identification of the stimuli up to the point where the consonant joined the vowel 82.8 percent of the time; this is highly significant ($x^2 = 92.03$, 1 df, $p < .001$). Conversely, the same statistic shows that the subjects' choice of the incorrect CVN responses were much below chance level. Lahiri and Marslen-Wilson state that according to the notion that redundant features are specified in the lexicon and accessed by listeners when interpreting the incoming speech signal (i.e., the hypothesis contrary to underspecification theory) “listeners should *never* interpret CVCs as potential CVNs” [italics added]. But this is an unrealistic requirement. Rather than “never”, the most that can be required is that they give CVN responses to gated CVC stimuli much less than would be predicted by chance; this is what Ohala and Ohala found.

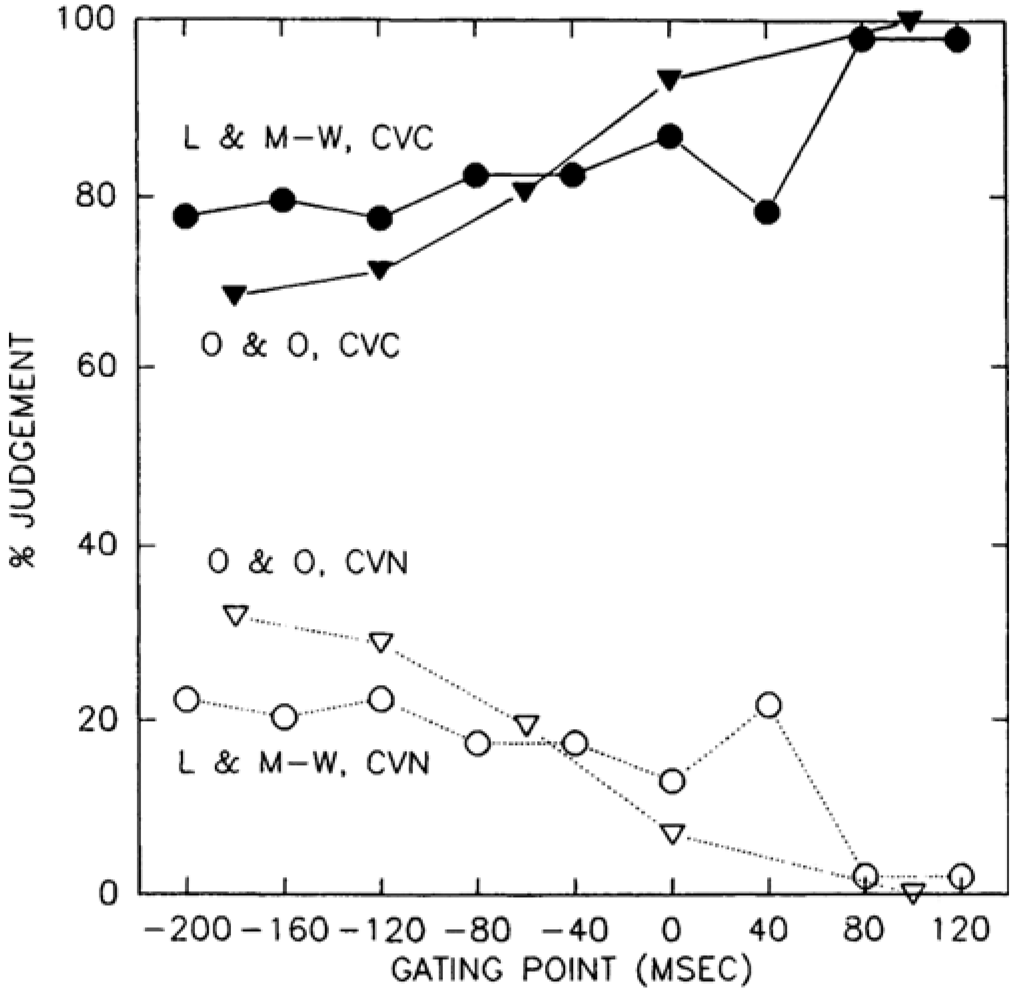

Figure 24.1 Listeners's judgements (vertical axis) of CVC syllables truncated at various points (horizontal axis, gating point in msec, where 0 = the VC boundary): Circles: results of Lahiri and Marslen-Wilson (1992); triangles: result of Ohala and Ohala (1993). Solid lines: CVC judgements; dotted line: CVN judgements. See text.

2 Do Experiments Ever Settle Issues?

Leaving aside divine revelation, there are no perfect routes to the truth (if one believes truth exists) and experiments are no exception. Being performed by fallible humans, they can be fallible, too. The answer to an experiment suspected of being flawed is a better-controlled experiment which overcomes the flaw. Thus experimental phonology or experiment anything should be viewed as a spiral process: make a claim; test the claim; revise (or abandon) the claim; test the revised claim, etc. Ultimately, this continuous process should lead to a convergence of results which suport a more confidently held belief.

1 It probably goes without saying, but I'll say it anyway: that it is only reasonable to subject to experimental study those claims which matter - that is, where, if they are wrong, there are major consequences for the way people think and act - and those which are connected in some way, even if indirectly, with the observable universe. Claims that a given phenomenon may be described, labeled, or classified in a certain way are candidates for debate rather than experimentation, unless the descriptions are grounded in empirical properties. Claims that don't specify the domain of the universe to which they apply - physical, psychological, social - would be hard to test experimentally. But equally, if one cannot specify the domain in which the claim holds, how did the belief come about in the first place? A belief arises from some constellation of evidence, even if suspect; the domain from which this evidence came is the domain in which the claim is to be tested.

2 The “uniformitarian hypothesis” guided and catalyzed the development of scientific geology in the 18th century. Rather than assuming cataclysmic events, often thought to have divine origins, it was posited that for the most part constant processes like erosion, sedimentation, and land upheaval accounted for the formation of mountains, valleys, and other conspicuous features of the earth's surface.

3 Lahiri and Marslen-Wilson make asymmetrical claims about how listeners will judge CVN and CVC stimuli with the final consonants gated off. Although both syllable types are said to have vowels which are lexically unspecified for the feature nasal, only the CV(C) stimuli are claimed to be ambiguous; the CV(N) stimuli they allow will be unambiguously identifiable as CVN because of the post-lexical rule spreading [+nasal] from the N to the preceding V. This admission strikes me as undercutting the claim, quoted above, that “surface structures derived after postlexical spreading do not play a distinctive role in perception.” In fact, there is no major dispute that listeners' *can* identify CVN syllables with the final N gated off (Ali, Gallagher, Goldstein, and Daniloff 1971). The present discussion is concerned with how to interpret listeners' reaction to end-truncated CVC stimuli.

4 The caption to the Lahiri and Marslen-Wilson (1992) figure (fig. 9.4) on which figure 24.1 was based indicates that the curves show percentage of response types. However, the abscissa in this figure clearly does not correspond to percentages and in fact is labeled “mean number of responses.” I have tried to make a reasonable interpretation of the data they presented, but the data in figure 24.1 could be off by 2-3 percent.

5 Lahiri and Marslen-Wilson (1991, 1992) also studied the role of vowel nasalization and other phonological phenomena in Bengali; Ohala and Ohala (1993) similarly reported a study of vowel nasalization in Hindi. I report here only the results pertaining to English.

Cite this article

OHALA, JOHN J. "Experimental Phonology." *The Handbook of Phonological Theory*. Glodsmith, John A. Blackwell Publishing, 1996. Blackwell Reference Online. 31 December 2007 <http://www.blackwellreference.com/subscriber/tocnode?id=g9780631201267_chunk_g978063120126726>

Bibliographic Details

The Handbook of Phonological Theory

Edited by: John A. Glodsmith
eISBN: 9780631201267
Print publication date: 1996