JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR **9**, 295–302 (1970)

The Nonperceptual Reality of the Phoneme[1]

H. B. SAVIN AND T. G. BEVER

University of Pennsylvania, Philadelphia, Pennsylvania 19104, and The Rockefeller University, New York, New York 10021

Subjects responded as soon as they heard a preset target in a sequence of nonsense syllables. The target was a complete syllable (e.g., "baeb" "saeb") or a phoneme from that syllable, the syllable-initial consonant phoneme for some objects (e.g., "b-" or "s-"), and the medial vowel phoneme for other subjects (e.g., "-ae-"). Subjects responded more slowly to phoneme targets than to syllable targets (by 40 msec for /s-/, 70 msec for /b-/ and 250 msec for medial /ae/). These results indicate that phoneme identification is subsequent to the perception of larger phonological units. The reality of the phoneme is demonstrated independently of speech perception and production by the natural presence of alphabets, rhymes, spoonerisms, and interphonemic contextual constraints.

A speech utterance can be segmented at different linguistic levels into phrases, words, syllables, and phonemic and phonetic segments. There is general agreement among linguists about what these levels are and how to represent any given utterance on each of the levels. Current theories of speech perception assume that, in the course of perceiving an utterance, listeners analyze the utterance at these linguistic levels. At present, however, little is known about the sequence of perceptual steps which transform the acoustic input into this multileveled perceptual representation. Almost the only thing one can say with certainty is that the sequence is not always "bottom-to-top": people do not always recognize phonetic segments first, then phonemic ones, then syllables, and so on. The *prima facie* evidence for this is that people use high-level (syntactic or semantic) context to resolve phonetic ambiguities in everyday speech or in speech with experimentally introduced background noise. For example, if *cat* and *hat* are presented out of context in a high background noise they are confused with one another, but in the context: *it's time to feed the . . .*, everyone hears *cat*, whether right or wrong.

The present report concerns the order in which listeners make decisions at the phonemic and syllabic levels in the course of speech perception. The method is to ask a listener to monitor a sequence of nonsense syllables for the presence of a certain linguistic unit, either a phoneme or a syllable, and to respond as quickly as possible when he has heard it. The results show that subjects respond consistently faster when given syllable targets to listen for than when given phoneme targets.

EXPERIMENT I

Method

Materials and apparatus. Fifty sequences of nonsense syllables which follow the phonological laws of English were recorded at the rate of one syllable per second ("Tape 1"). (The syllables were a mean of .6 sec in duration.) The sequences ranged in length from 5 to 17 different syllables. In each sequence there was exactly one target syllable beginning with /b/, occurring at a point between the 3rd syllable and 13th syllable, with roughly equal probability for each position in the sequence. Each target syllable beginning in /b/ appeared in only one list and was followed by 2–4 other syllables. Sample sequences are presented in Table 1. (Out of

[1] This research was supported by the National Institutes of Health Grant No. 1 POL GM 16735–01; the Advanced Research Projects Agency Contract No. DAHC 15 to The Rockefeller University; and NSF GB 6529 to The University of Pennsylvania. We are indebted to G. A. Miller for stimulating the research, to R. Hurtig, R. Littky, and B. Schemmel for assistance in the experimentation and V. Valian for advice on this manuscript.

TABLE 1

SAMPLE SEQUENCES USED IN EXPERIMENTS I, II, III[a]

				Experiment I				
thowj	tuwp	barg	tiyj	friyn				
skaem	firt	mowf	kiyg	gib	nemp	bolf	dayts	saymz
haen	stot	hawt	reg	thim	gowj	telg	mub	forg
nuwj	tayn	bruwj	ins	stowg	welb			
				Experiment II				
thowj	tūwp	sarg	tiyf	friyn				
kaemf	firt	mowf	kiyg	gib	nepm	solf	dayt	baym
haen	torb	hawt	reg	bim	gowj	telg	mub	forg
nuwj	tayn	sluwj	int	towg	welb			
				Experiment III				
thowj	tuwp	baelg	tiyf	friyn				
skem	firt	muwf	kigg	gib	nemp	baelf	dayt	saymz
hin	stat	hawt	vey	thim	gowj	telg	mub	forg
muwj	tayn	braej	ins	stowg	welb			

[a]The targets are underlined.

407 unique non target syllables (i.e., those with initial consonants other than /b/) 15 appeared in two separate sequences. The distribution of sequences of different length was roughly balanced throughout the 50 trials.) No syllables in these sequences began in /p/ or /v/, because of the high confusability of these sounds with initial /b/.

The speech was recorded on one track of a two-track tape recorder. On the second track of the tape a tone was recorded, starting approximately at the beginning of the /b/ syllable and continuing for several seconds. This tone was used to close a relay which was in series with an electronic timer and a closed telegraph key. The timer started when the tone operated the relay and stopped when *S* released the telegraph key (*S* could not hear the timing tone). We located the beginning of the target by moving the tape manually back and forth across the playback head. There is undoubtedly some error in the alignment of tone onset with the beginning of the target syllable, and there is also a constant error of about 40 msec because the relay does not close instantaneously. However, because of the counterbalanced procedure (every trial was a phoneme target trial for half the subjects and a syllable target trial for the other half), any timing error would have the same effect on the average reaction time for phoneme targets and for syllable targets.

Procedure. The *S* was trained on the reaction time task by responding to the onset of a 500 Hz tone for six trials. Each *S* then heard the 50 sequences one by one, played over earphones at a comfortable listening level. The *S* was given general typewritten instructions outlining the experiment. In half the trials ("phoneme target trials") *S* was told that he should respond to the first syllable beginning in /b/ without knowing what the the rest of the syllable would be. In the other half of the trials ("syllable target trials") *S* was told he would first hear the entire target syllable and be given a written quasi-English spelling of the target syllable (e.g., "bab" for the target /baeb/). Each trial was then introduced by an auditory presentation of the trial number 2 sec before the actual sequence (e.g., "trial no. 15"). Before every other trial *S* heard the target syllable and could read it on the sheet in front of him (e.g., "The target is baeb"). On the other trials *S* heard no target, and knew that his target was the first syllable with initial /b/. For half the subjects the syllable targets were presented on the odd-numbered trials and for the other half of the subjects the syllable targets were presented on the even-numbered trials. Each session lasted for about 30 min.

Subjects. Subjects were 29 college undergraduates from the New York City area who volunteered their time for paid participation (at $2.50/hr.). Fifteen *S*s were offered a $2.00 bonus if "the average of all their responses was faster than that of the average subject" (condition IA). Eight *S*s were offered "a $.25 bonus for each response faster than 150 msec" (condition IB). Six *S*s were offered "a $.25 bonus for each trial when their reponses on phoneme target trials were faster then 150 msec," but were given no extra motivation for syllable target trials (condition IC).

TABLE 2

MEAN REACTION TIME (SEC) TO SYLLABLE AND PHONEME TARGETS

	Condition	Phoneme target	Syllable target	Difference	S_M of difference
Exp. I initial /b-/	A	.333	.259	.074	.006
	B	.314	.195	.119	.024
	C	.306	.234	.072	.023
Exp. II initial /s-/		.354	.311	.043	.010
Exp. III medial vowel /-ae-/		.470	.225	.245	.034

Results

The results are displayed in Table 2-I. All *S*s in all payoff conditions responded faster on syllable target trials than on phoneme target trials ($p < .001$, sign test across *S*s). The false alarm rate was less than 3% for both syllable and phoneme targets with no noticeable difference in false alarm rates for the two kinds of targets.

Reaction times improved during the experimental session, as shown in Figure 1 for the 15 *S*s in condition A, but the phoneme/syllable target difference remained remarkably constant (86 msec for the first 25 trials, 64 msec for the last 25 trials).

The low false alarm rate for all *S*s in all payoff conditions makes it difficult to examine the possible role of a shifting decision criterion in the relatively long reaction to phoneme

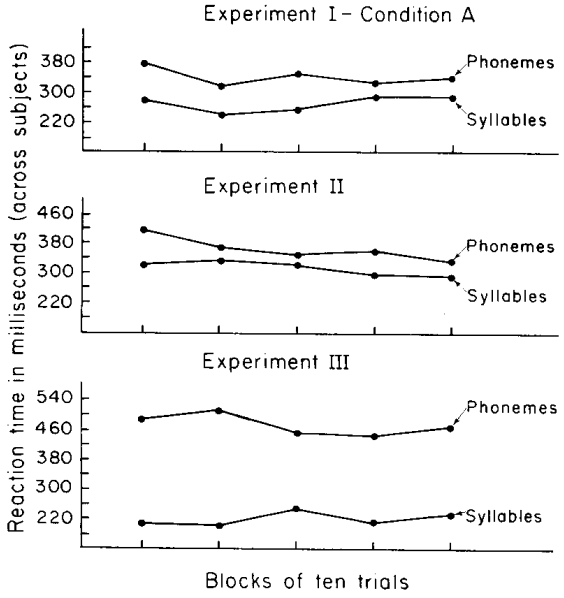

FIG. 1. Mean reaction time (sec) to phoneme and syllable targets and the mean difference during the experimental session for the 15 *S*s in condition A of Experiment I, 17 *S*s in Experiment II, and 6 *S*s in Experiment III.

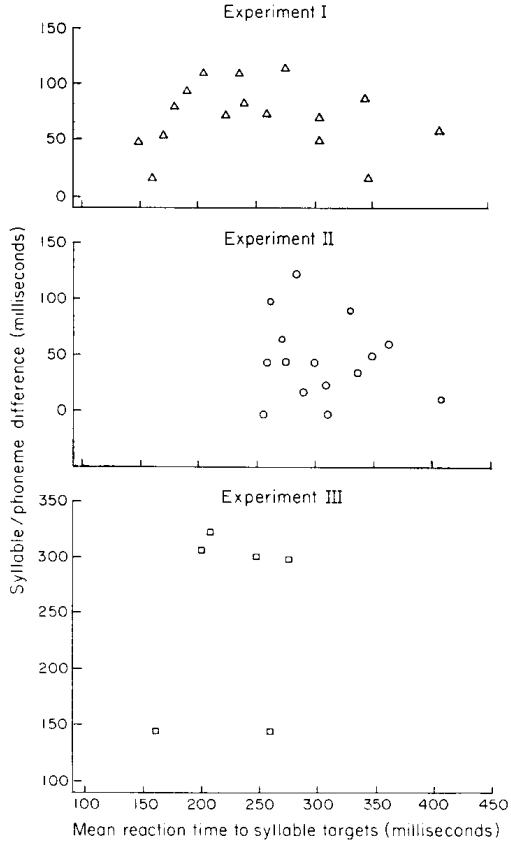

FIG. 2. Relationship between an *S*'s mean syllable reation time and his mean phoneme/syllable reaction time difference. Each point represents one *S*. (The 15 *S*s in Experiment I are those in condition A; see Table 1.)

targets. Such variables appear to have little effect here since the phoneme/syllable difference occurs even in condition IC in which the payoffs were explicitly designed to encourage fast responding to phoneme targets but not syllable targets. Figure 2 presents the phoneme/syllable target difference for *S* as a function of *S*s' reaction time to syllable targets. The result displayed in Figure 2-I suggest that the discrimination of the initial /b-/ phoneme adds a constant amount of time to the recognition of the syllable.

Experiment II

The acoustic shape of an initial stop consonant like /b/ differs greatly depending on the vowel which follows (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Thus, the finding in Experiment I that reaction time is faster if *S*s are given syllable targets might be due to their knowing just which variant of initial /b/ they were listening for. To test this hypothesis we examined *S*s' responses to syllables beginning in /s/, since the acoustic shape of /s/ is relatively unaffected by the following vowel.

Method

The recorded materials, apparatus, and procedure were identical with those in condition A of Experiment I except that the critical syllable began in /s/ rather than /b/ ("Tape 2"). Also there were no syllables which began in /z/ or /š/ because of the confusability of these sounds with initial /s/. Seventeen *S*s were offered $2.00 extra for relatively fast average responses.

Results

The mean reaction time to /s/-syllables was 43 msec slower when *S*s were given the phoneme /s/ as a target than when given the syllable as a target (Table 2, Experiment II). This difference is significant ($p < .01$, sign test across *S*s) but is smaller than that found for the /b/ syllables ($p < .01$ by t test on the differences between syllables and phoneme targets). The effects of the experimental session were the same as in experiment I (Figure 1-II). (The mean phoneme/syllable target differences was 47 msec for the first 25 trials and 40 msec for the second 25 trials.) Furthermore, as in Experiment I, the syllable/phoneme target difference is quite constant regardless of *S*'s mean rate on syllable targets (Figure 2-II). (It should be noted that the *absolute* differences in recorded reaction time between /s-/ syllables and /b-/ syllables might be due to systematic differences in placing the timing tone at the beginning of a /b-/ syllable as opposed to an /s-/ syllable.)

Experiment III

The fact that being told the entire syllable facilitates reaction time less if the syllable begins in /s/ than if it begins in /b/ might be taken as evidence that the difficulty of responding to an initial consonant alone is due to the difficulty of discriminating a consonant without knowledge of the following vowel. On this view *S* waits to identify the vowel before deciding that the initial consonant is (or is not) the target: if the target is given as /b/ the *S* must identify more of the vowel and choose from more acoustic possibilities than if the consonant is /s/, since the acoustic shape of /b/ is more affected by the following vowel than /s/.

This interpretation of the results could not explain the fact that responses to /s/ targets are slower than they are to the corresponding syllable targets, since every acoustic version of /s/ shares a unique acoustic property not shared by any other initial consonant (in our materials): a burst of high-frequency acoustic energy. However, to test further the view that *S*s first identify the vowel and then the preceding consonant, we studied their reaction time to syllables in which the target was the medial vowel. If vowels are indeed identified as part of the process of identifying the preceding consonants, then vowel targets should be responded to faster than initial consonant targets. To put it another way, if the difference in reaction times in the preceding experiments was critically due to lack of knowledge of the vowel, then being told the vowel alone should

lead to responses as fast as being told the entire syllable.

Method

Materials and apparatus. A new tape of 50 trials was recorded ("Tape 3"). This tape was identical with the ones in the previous experiments except that the syllable in the target position always had the vowel /ae/ (as in "cat"), with different initial and final consonants. Also, no prior syllable had the vowel /a/ (as in "car") because of its confusability with /ae/.

Subjects and procedure. The *S*s were pretrained and instructed as in previous experiments. Six *S*s listened to Tape 3 and heard the entire target syllable or were to respond to the first syllable containing the medial vowel /ae/ on alternate trials. The *S*s were offered an extra $2.00 for relatively fast average responses.

Results

The *S*s responded more slowly to vowel targets than to syllable targets. Furthermore, the magnitude of the difference between syllable targets and phoneme targets is larger when the phoneme target is a medial vowel than when it is an initial /b/ ($p < .01$ t test for the pooled results of all conditions in Experiment I compared with Experiment III). Practice effects were slight, as in Experiments I and II (Figure 1-III). (The difference is 277 msec for the first 25 trials and 213 msec for the last 25 trials for *S*s in experiment III.) The syllable/phoneme reaction time difference was roughly constant, independent of *S*'s syllable target reaction time (Figure 2-III).

Three Morals for the Description of Language

The Perceptual Moral

The results of Experiment III refute the view that phoneme targets are responded to more slowly than syllable targets because listeners must identify the vowel before identifying the preceding phoneme. Experiment II demonstrated that the relative slowness of reaction times to initial consonant phonemes is not due to the fact that there are several acoustically distinct allophones which reflect prevocalic consonants. Thus, phonemes are not perceived before the syllabic units that contain them (at least not before initial conconant-vowel sequences).

An alternative view of speech perception is that syllables are perceived before their constituent phonemes. The simplest variant of this hypothesis would predict that once the syllable is perceived any phoneme would be identifiable as quickly as any other. On this view, the phoneme/syllable identification time difference should be as great for initial consonant targets as for vowel targets. However, the results in Experiment III show that this is not the case; rather, vowel targets take much longer relative to syllable targets than do initial consonant targets. The relative size of the effect for vowels could be due either to a "left-to-right" phoneme identification process or to one which gives precedence to the identification of all consonants regardless of their order (or some other basis for ordering). Whichever of these possibilities is correct, the studies in this paper support the view that phonemes are identified only after some larger linguistic sequence (e.g., syllables or words) of which they are parts.

The Methodological Moral

Foss and Lynch (1969) and Foss (1969) recently measured reaction times to word-initial phoneme targets in English sentences and found that the reaction times depend upon what precedes the target phoneme: responses to a phoneme target are slower when the words they begin followed an unpredictable or syntactically complex sequence. Foss (1969) interpreted these results as consistent with the view that "as decision difficulty concerning the identification of entities at one level increases, speed and accuracy of decisions concerning other entities should decrease since the difficult decisions utilize the bulk of *S*s limited-capacity mechanisms [p. 457]". In other words, the phonemes that occur in syntactically "difficult" or unpredictable contexts are perceived slowly not because of any special connection between phoneme recognition and syntactic and predictive analysis but

simply because, like almost any other two sorts of decision making, they compete with one another for the same nonspecific "central analyzing mechanisms". He concludes that ". . . comprehension models . . . that propose non-overlapping analyzing mechanisms for the identification of phonological and lexical items seem to be ruled out" by the phoneme detection results [p. 462].

A general-purpose central analyzing mechanism that can make phonemic, lexical, and syntactic decisions, but not all at the same time, may exist, but the phoneme discrimination data are not persuasive evidence for it. If, as demonstrated in our experiments, listeners do not identify phonemes except in already-identified syllables, then anything that makes syllable recognition faster will *ipso facto* increase the speed of the recognition of the constituent phonemes. Being in a predictable context and being in a phrase that is easy to analyze syntactically makes the recognition of whole words relatively fast—indeed word recognition is a necessary part of syntactic and probabilistic analysis so it is to be expected that complex syntactic structures lead to slow identification of the following words. Syllables within slowly recognized words are themselves slowly recognized, and, if syllable recognition invariably precedes phoneme recognition, then syntactic complexity can affect phoneme identification via effects on word recognition. That is, the major variable which affects differences in "phoneme detection" time is variation in syllable (or word) detection time. This explanation does not postulate any attention-distribution "central analyzer" nor does it require that phonetic and syntactic perceptual processes compete directly for attention.

The Polemical Moral

Phonemes (and phonemic features) have been awkward concepts to manage for those psychologists and linguists who wish theoretical entities to be close to the acoustic or physiological periphery. We take our results to show that phonemes are perceived only by an analysis of already perceived syllables (or at least already perceived consonant-vowel pairs). The question then arises, what is the point of having a system of speech perception that can carry out phonemic analysis? Having perceived the syllable, why not always proceed directly to morphemes, phrases, and other semantically relevant units? This question immediately raises the further one: what is the evidence that people normally *do* (outside of experiments like ours and grade-school reading classes) bother with the seemingly superfluous analysis of syllables into phonemes?

Some scholars might conclude from our results that phonemes must be principally relevant to the *production* of speech since we have found them to be incidental to perception; the argument would be that if phonemes are not perceptual then the only thing left for them to be is motor. However, the fact that coarticulation effects can modify the motor behavior in the production of the same phoneme invalidates any purely peripheral notion of behavioral response as a basis for a definition of phonemes. Indeed, Wickelgren (1969) has recently invoked the facts about coarticulation as part of an attempt to disprove the existence of phonemes as psychological entities of any sort. For example, he denies that there is any psychological unit that occurs twice in the word *did* since the initial and the final /d/ are different articulatorily and acoustically. They are similar enough so that some careless people *say* they are the same, but they are not similar by virtue of being the physical realizations in different contexts of the same abstract psychological entity, namely, the phoneme /d/ (whose existence Wickelgren denies). A thorough discussion of Wickelgren's views is beyond the scope of this paper. Ultimately, he makes it clear that he must deny the existence of phonemes not because of the facts he cites but because the existence of phonemes is incompatible with his general beliefs about the nature of associations and their central place in psychological theory.

What he claims for the behavioral facts about phoneme production is that they do not force him to believe in phonemes—not that they force him to disbelieve in them. It is this line of negative argument that makes his work relevant to our present paper. We find ourselves in the same position with respect to the perceptual facts; they do not seem to require phonemes for their explanation (except in one respect, whose theoretical import is quite unclear: the subjects did somehow know what we were talking about when we gave them phoneme-target instructions).

Are phonemes then dispensable as psychological constructs, since neither the perceptual nor the articulatory facts require them? Not at all. It is quite impossible to do without phonemes in psychological theories of language, but for nonsensory, nonarticulatory reasons that illuminate both the nature of phonemes and the nature of the relationship between speech behavior and linguistic structure. To mention only a few aspects of the behavioral evidence for phonemes, there is the occurrence of alphabetic writing systems, the existence of rhyme and alliteration in nonliterate poetry, the natural existence of segmental phonemic spoonerisms, and the innumerable well-attested historical changes in language that are described very simply in terms of phonemes and only clumsily and arbitrarily without them. (More accurately, these descriptions are in terms of phonemic features, but the difference does not affect the present argument since many features are necessarily internal to phonemic segments.)

In addition to such historical facts (which could hardly be facts but for some psychological facts about the people who make the history) there are numerous regularities in every modern language which can be stated satisfactorily only by referring to phonemic features. Consider, for example, just part of the rule of plural formation for modern English nouns.

If the (singular) noun ends in /s/ or /z/, the plural is the syllable /ez/ (horses, roses); otherwise, if it ends in a voiced sound, add /z/ (boys) and if it ends in an unvoiced sound add /s/ (bits).

To state such a rule in terms of unsegmented syllables would be a great deal more complex since all the syllables of each kind would have to be listed. Not only would this be inelegant, it would not represent the generalization which underlies the groupings of the three different kinds of syllables. Segmental phonemic features turn out to be just the appropriate concepts for such phenomena (Chomsky & Halle, 1968)—neither too specific nor too undifferentiated to describe the sorts of regularities like this that keep turning up. Again, how could this be if those phonemic features were not part of a psychologically correct description of what people know intuitively about the sound structure of their language?

The conclusion that follows from such considerations is that phonemes are primarily neither perceptual nor articulatory entities. Rather, they are psychological entities of a nonsensory, nonmotor kind, related by complex rules to stimuli and to articulatory movements, but they are not a unique part of either system of directly observable speech processes. In short, phonemes are *abstract*. The adaptive value of such abstract entities remains to be clarified. Just by virtue of standing neutral between sensory input and articulatory output, they can interrelate the perceptual and expressive speech processes. This is most certainly a useful and crucial function for phonemes. But the problem remains to understand why this function is served with entities that have so complex a relationship both to acoustic stimuli and to articulatory movements.

References

Chomsky, N., & Halle, M. *The Sound Patterns of English*. New York: Harper & Row, 1968.

Foss, D. J., & Lynch, R. H. Decision processes during sentence comprehension: Effects of surface structure on decision times. *Perception and Psychophysics*, 1969, **5**, 145–148.

Foss, D. J. Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behavior*, 1969, **8**, 457–462.

LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. Perception of the speech code. *Psychological Review*, 1967, **74**, 431–461.

WICKELGREN, W. A. Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, 1969, **76**, 1–15.

(Received November 19, 1969)