Psychology and the segment

ANNE CUTLER

Something very like the segment must be involved in the mental operations by which human language users speak and understand.* Both processes-production and perception - involve translation between stored mental representations and peripheral processes. The stored representations must be both *abstract* and *discrete.*

The necessity for abstractness arises from the extreme variability to which speech signals are subject, combined with the finite storage capacity of human memory systems. The problem is perhaps worst on the perceiver's side; it is no exaggeration to say that even two productions of the same utterance by the same speaker speaking on the same occasion at the same rate will not be completely identical. And within-speaker variability is tiny compared to the enormous variability across speakers and across situations. Speakers differ widely in the length and shape of their vocal tracts, as a function of age, sex, and other physical characteristics; productions of a given sound by a large adult male and by a small child have little in common. Situation-specific variations include the speaker's current physiological state; the voice can change when the speaker is tired, for instance, or as a result of temporary changes in vocal-tract shape such as a swollen or anaesthetized mouth, a pipe clenched between the teeth, or a mouthful of food. Other situational variables include distance between speaker and hearer, intervening barriers, and background noise. On top of this there is also the variability due to speakers' accents or dialects; and finally, yet more variability arises due to speech style, or register, and (often related to this) speech rate.

But the variability problem also exists in speech production; we all vary our speech style and rate, we can choose to whisper or to shout, and the

*This paper was prepared as an overall commentary on the contributions dealing with segmental representation and assimilation, and was presented in that form at the conference.

accomplished actors among us can mimic accents and dialects and even vocal-tract parameters which are not our own. All such variation means that the peripheral processes of articulation stand in a many-to-one relationship to what is uttered in just the same way as the peripheral processes of perception do.

If the lexicon were to store an exact acoustic and articulatory representation for every possible form in which a given lexical unit might be heard or spoken, it would need infinite storage capacity. But our brains simply do not have infinite storage capacity. It is clear, therefore, that the memory representations of language which we engage when we hear or produce speech must be in a relatively abstract (or normalized) form.

The necessity for discreteness also arises from the finite storage capacity of our processing systems. Quite apart from the infinite range of situational and speaker-related variables affecting how an utterance is spoken, the set of potential complete utterances themselves is also infinite. A lexicon-that is, the stored set of meaning representations-just cannot include every utterance a language user might some day speak or hear; what is in the lexicon must be discrete units which are smaller than whole utterances. Roughly, but not necessarily exactly, lexical representations will be equivalent to words. Speech production and perception involve a process of translation between these lexical units and the peripheral input and output representations. Whether this process of translation in turn involves a level of representation in terms of discrete sublexical units is an issue which psycholinguists have long debated.

Arguments in favor of sublexical representations have been made on the basis of evidence both from perception and from production. In speech perception, it is primarily the problem *of segmentation* which has motivated the argument that prelexical classification of speech signals into some subword-level representation would be advantageous. Understanding a spoken utterance requires locating in the lexicon the individual discrete lexical units which make up the utterance, but the boundaries between such units - i.e. the boundaries between words-are not reliably signaled in most utterances; continuous speech is just that-continuous. There is no doubt that a sublexical representation would help with this problem, because, instead of being faced with an infinity of points at which a new word might potentially commence, a recognizer can deal with a string of discrete units which offer the possibility of a new word beginning only at those points where a new member of this set of sublexical units begins.

Secondly, arguments from speech perception have pointed out that the greatest advantage of a sublexical representation is that the set of potential units can be very much smaller than the set of units in the lexicon. However large and heterogeneous the lexical stock (and adult vocabularies run into many tens if not hundreds of thousand items), with sublexical representations any lexical item could be decomposed into a selection from a small and finite set of units. Since a translation process between the lexicon and more peripheral processes is necessary in any case, translation into a small set *of* possibilities will be far easier than translation into a large set of possibilities.

Exactly similar arguments have been made for the speech-production process. If all the words in the lexicon can be thought of as being made up of a finite number of building blocks in various permutations, then the translation process from lexical representation to the representation for articulation need only know about how the members of the set of building blocks get articulated, not how all the thousands of entries in the lexicon are spoken.

Obvious though these motivating arguments seem, the point they lead to is far from obvious. Disagreement begins when we attempt to answer the next question: what is the nature of the building blocks, i.e. the units of sublexical representation? With excellent reason, the most obvious candidates for the building-block role have been the units of analysis used by linguists. The phoneme has been the most popular choice because (by definition) it is the smallest unit into which speech can be sequentially decomposed. I wish that it were possible to say at this point: the psycholinguistic evidence relating to the segment is unequivocal. Unsurprisingly, however, equivocality reigns here as much as it does in phonology.

On the one hand, language users undoubtedly have the ability to manipulate speech at the segmental level, and some researchers have used such performance data as evidence that the phoneme is a level of representation in speech processing. For instance, language games such as Pig Latin frequently involve movement of phoneme-sized units within an utterance (so that in one version of the game, *pig latin* becomes *ig-pay atin-lay).* At a less conscious level, slips of the tongue similarly involve movement of phoneme-sized units - "by far the largest percentage of speech errors of all kinds" says Fromkin (1971: 30), involve units of this size: substitution, exchange, anticipation, perseveration, omission, addition-all occur more often with single phonemes than with any other linguistic unit. The on-line study of speech recognition has made great use of the phoneme-monitoring task devised by Foss (1969), which requires listeners to monitor speech for a designated target phoneme and press a response key as soon as the target is detected; listeners have no problem performing this task (although as a caveat it should be pointed out that the task has been commonly used only with listeners who are literate in a language with alphabetic orthography). Foss himself has provided (Foss and Blank 1980; Foss, Harwood and Blank 1980; Foss and Gernsbacher 1983) the strongest recent statements in favor of the phoneme as a unit of representation in speech processing: "the speech perception mechanisms compute a representation of the input in terms of phoneme-sized units" (Foss, Harwood, and Blank 1980: 185). This argument is based on the fact that phoneme targets can be detected in heard speech prior to contact with lexical representations, as evidenced by the absence of frequency effects and other lexical influences in certain types of phoneme-monitoring experiment.

Doubt was for a time cast on the validity of phoneme-monitoring performance as evidence for phonemic representations in processing, because it was reported that listeners can detect syllable-sized targets faster than phoneme-sized targets in the same speech material (for English, Savin and Bever 1970; Foss and Swinney, 1973; Mills 1980; Swinney and Prather 1980; for French, Segui, Frauenfelder, and Mehler 1981). However, Norris and Cutler (1988) noted that most studies comparing phoneme- and syllable-monitoring speed had inadvertently allowed the syllable-detection task to be easier than the phoneme-detection task-when the target was a syllable, no nontarget items were presented which were highly similar to the target, but when the target was a phoneme, nontarget items with very similar phonemes did occur. To take an example from Foss and Swinney's (1973) experiment, the list *giant valley amoral country private middle bitter protect extra* was presented for syllable monitoring with the target *bit-*, and for phoneme monitoring with the target *b-*. The list contains no nontarget items such as *pitcher*, *battle*, or *bicker* which would be very similar to the syllable target, but it does contain nontarget items beginning with p-, which is very similar to the phoneme target. In effect, this design flaw allowed listeners to respond in the syllable-target case on the basis of only partial analysis of the input. Norris and Cutler found that when the presence of items similar to targets of either kind was controlled, so that listeners had to perform both phoneme detection and syllable detection on the basis of equally complete analyses of the input, phoneme-detection times were faster than syllable-detection times. Thus there is a substantial body of opinion favoring phonemic units as processing units.

On the other hand, there are other psycholinguists who have been reluctant even to consider the phoneme as a candidate for a unit of sublexical representation in production and perception because of the variability problem. To what degree can it be said that acoustic cues to phonemes possess constant, invariant properties which are necessarily present whenever the phoneme is uttered? If there are no such invariant cues, they have argued, how can a phonemic segmentation process possibly contribute to processing efficiency, since surely it would simply prove enormously difficult in its own right? Moreover, at the phoneme level the variability discussed above is further compounded by coarticulation, which makes a phoneme's spoken form sensitive to the surrounding phonetic context - and the context in question is not limited to immediately adjacent segments, but can include several segments both forwards and backwards in the utterance. This has all added up to what seemed to some researchers like insuperable problems for phonemic representations.

As alternatives, both units above the phonemic level, such as syllables, demisyllables, or diphones, and those below it, such as featural representations or spectral templates, have been proposed. (In general, though, nonlinguistic units such as diphones or demisyllables have only been proposed by researchers who are concerned more with machine implementation than with psychological modeling. An exception is Samuel's [1989] recent defense *of* the demisyllable in speech perception.) The most popular alternative unit has been the syllable (Huggins 1964; Massaro 1972; Mehler 1981; Segui 1984), and there is a good deal of experimental evidence in its favor. Moreover, this evidence is very similar to the evidence which apparently favors the phoneme; thus language games often use syllabically defined rules (Scherzer 1982), slips of the tongue are sensitive to syllabic constraints (MacKay 1972), and on-line studies of speech recognition have shown that listeners can divide speech into syllables (Mehler *et al.* 1981).

Thus there is no unanimity at all in the psycholinguistic literature, with some researchers favoring phonemic representations and some syllabic representations, while others (e.g. Crompton 1982) favor a combination of both, and yet others (e.g. Samuel 1989) opt for some more esoteric alternative. A consensus may be reached only upon the lack of consensus. Recent developments in the field, moreover, have served only to sow further confusion. It turns out that intermediate levels *of* representation in speech perception can be language-specific, as has been shown by experiments following up the finding of Mehler *et al.* (1981) that listeners divide speech up into syllables as they hear it. Mehler *et al.'s* study was carried out in French; in English, as Cutler *et al.* (1986) subsequently found, its results proved unreplicable. Cutler *et al.* pointed out that syllable boundaries are relatively clearer in French than in English, and that this difference would make it inherently more likely that using the syllable as a sublexical representation would work better in French than in English. However, they discovered in further experiments that English listeners could not divide speech up into syllables even when they were listening to French, which apparently encourages such a division; while French listeners even divided English up into syllables wherever they could, despite the fact that the English language fails to encourage such division. Thus it appears that the French listeners, having grown up with a language which encourages syllabic segmentation, learnt to use the syllable as an intermediate representation, whereas English listeners, who had grown up with a hard-to-syllabify language, had learnt not to use it.

In other words, speakers' use of intermediate units of representation is determined by their native language.

The reason that this finding muddied the theoretical waters is, of course, that it means that the human language-processing system may have available to it a range of sublexical units of representation. In such a case, there can be no warrant for claims that any one candidate sublexical representation is more basic, more "natural" for the human language-processing system, than any other for which comparable evidence may be found.

What relevance does this have to phonology (in general, and laboratory phonology in particular)? Rather little, perhaps, and that entirely negative: it suggests, if anything, that the psychological literature is not going to assist at all in providing an answer to the question of the segment's theoretical status in phonology.

This orthogonality of existing psycholinguistic research to phonological issues should not, in fact, be surprising. Psychology has concluded that while the units of sublexical representation in language perception and production must in terms of abstractness and discreteness resemble the segment, they may be many and varied in nature, and may differ from language community to language community, and this leaves phonology with no advance at all as far as the theoretical status *of* the segment is concerned. But a psychological laboratory is not, after all, the place to look for answers to phonological questions. As I have argued elsewhere (Cutler 1987), an experiment can only properly answer the question it is designed to answer, so studies designed to answer questions about sublexical representations in speech processing are *ipso facto* unlikely to provide answers of relevance to phonology. When the question is phonological, it is in the phonology laboratory that the answer is more likely to be found.