JOURNAL OF MEMORY AND LANGUAGE 32, 258–278 (1993)

Mora or Syllable? Speech Segmentation in Japanese

TAKASHI OTAKE AND GIYOO HATANO

Dokkyo University, Tokyo, Japan

ANNE CUTLER

MRC Applied Psychology Unit, Cambridge, United Kingdom

AND

JACQUES MEHLER

Laboratoire de Sciences Cognitives et Psycholinguistique, Paris, France

Four experiments examined segmentation of spoken Japanese words by native and nonnative listeners. Previous studies suggested that language rhythm determines the segmentation unit most natural to native listeners: French has syllabic rhythm, and French listeners use the syllable in segmentation, while English has stress rhythm, and segmentation by English listeners is based on stress. The rhythm of Japanese is based on a subsyllabic unit, the mora. In the present experiments Japanese listeners' response patterns were consistent with moraic segmentation; acoustic artifacts could not have determined the results since nonnative (English and French) listeners showed different response patterns with the same materials. Predictions of a syllabic hypothesis were disconfirmed in the Japanese listeners' results; in contrast, French listeners showed a pattern of responses consistent with the syllabic hypothesis. The results provide further evidence that listeners' segmentation of spoken words relies on procedures determined by the characteristic phonology of their native language.

INTRODUCTION

shinshin-to
ume chiri-kakaru
niwabi kana
Kakei (School of Basho)

The haiku is a Japanese verse form, the structure of which is rigidly prescribed: a haiku consists of seventeen morae in three groups of five, seven, and five morae, respectively. Kakei's haiku, above, meets these formal constraints.

When Japanese poetic forms such as the haiku are rendered in other languages, an approximation to the prescribed form is usually achieved by specifying the number of syllables per line. Thus one could freely translate the above haiku as:

In silence they fall
Into the garden bonfire
Petals of the plum.

But morae do not necessarily correspond to syllables. Consider the first line of the haiku: *shinshin-to*. Although it has the prescribed 5 morae (*shi, n, shi, n, to*), it has only 3 syllables (*shin, shin, to*). The mora is a *subsyllabic* unit; it can be a vocalic nucleus, a nucleus plus syllable onset, or, as

This research was supported by a grant from the Human Frontier Scientific Program. We are very grateful for the significant contributions to this project made by Kazuhiko Kakehi, Dennis Norris, and Juan Segui, and for the experimental assistance provided by Kiyoko Yoneyama (in Tokyo), Sally Butterfield, Ruth Kearns, and Duncan Young (in Cambridge), and Anne Christophe, Emmanuel Dupoux, Caroline Floccia, Emmanuelle Le Louarn, Kazuaki Miyagishima, and Christophe Pallier (in Paris). For further useful suggestions and assistance we thank Mary Beckman, Karalyn Patterson, and Taeko Wydell.

Address reprint requests to Takashi Otake at Department of English, Dokkyo University, 1–1 Gakuen-machi, Soka-shi, Saitama 340, Japan.

0749-596X/93 $5.00

in the second and fourth morae of *shinshin-to,* it can be the postvocalic portion of a syllable, i.e., the coda. In Japanese, a language with a very restricted phonological inventory, there are 108 distinct morae, and they are of only 5 types: CV, CCV, V, nasal coda (which we can represent as N), and geminate (doubled) consonant (represented as Q). There is only one nasal coda mora (which represents the sounds [n], [m], or [ŋ] depending on the following phonetic context), only one geminate consonant mora (which represents a doubling of whichever consonant happens to follow it), and only 5 vowels. Some morae begin with affricated consonants and hence can be considered CCV; otherwise, the second consonant in CCV morae can only be the glide [j], and only eight consonants may precede and 3 vowels follow [j]. Thus over 60% of all possible morae are of the simple CV type.[1]

This obviously suggests that there will also be a skew towards CV morae in the frequency distribution of mora types in Japanese speech, and this is indeed so: just as 14 of the 17 morae in the haiku above are CV, so too do we find that in corpora of Japanese speech more than 70% of morae are CV (Otake, 1990). Thus in practice the mora will often overlap with the syllable. Indeed, many morae *are* syllables: the allowable syllable structures of Japanese include V, CV, and CCV, which are all 1-mora syllables. (The remaining allowable syllables—any of the above 3 structures followed by N, Q, or V—all contain 2 morae.)

However, it is the mora and not the syllable which is important for the specification of verse forms. Moreover, the mora also plays a central role in Japanese orthography. Japanese has a mixed orthographic system: content words are written in *kanji,* which are Chinese-based ideographic characters, while function words and inflectional affixes, plus foreign words, are written in *kana.* There are two *kana* orthographies, *hiragana* and *katakana,* and both are mora-based: each contains about 50 symbols, which alone or in combination can be used to represent each mora of the language.

The mora can be thought of properly as a rhythmic unit—the geminate consonant marker, for instance, implies consonantal lengthening. Thus Japanese has been said to be "mora-timed," in contrast to the "stress-timing" characteristic of English and other stress languages, or the "syllable-timing" characteristic of French and many other languages. Note that this does *not* mean that every mora in an utterance occupies exactly the same amount of time (Beckman, 1982), any more than stress units occur at fixed temporal intervals in English (Dauer, 1983), or syllables occur at fixed temporal intervals in French (Wenk & Wioland, 1982). Each of these units may vary in the number and type of phonetic segments it contains, and because some sounds simply take longer to articulate than others, these factors will influence the realized duration. Taking such factors into account, however, the mora certainly seems to be a strongly predictive factor in articulatory timing of Japanese speech (Port, Dalby & O'Dell, 1987; although note that a subset of Port et al.'s findings, for CV structures, are replicable in English and Spanish: Otake, 1989). Evidence from speech errors (Kubozono, 1989) and lan-

[1] This description is phonetically based, i.e., the number of morae is defined as the number of phonologically legal mora structures having distinct IPA transcriptions. Thus it does not correspond exactly to the number of characters in the *kana* orthographies. These allow a very limited degree of mismatch with a phonetic transcription. For instance, a phonetic transcription distinguishes between a velar nasal + vowel sequence and a velar stop + vowel sequence, while the *kana* orthographies do not; conversely, the orthographies contain a couple of pairs of characters with identical transcriptions, such as the two characters pronounced [dʑi]. Also, it should be noted that the *kana* orthographies encode some further underlying phonological relationships between morae; for example, two morae beginning with a consonant having the same manner and place of articulation but differing in voicing are represented by characters which differ only in the presence versus absence of a voicing marker.

guage games (Katada, 1990) further confirms a role for the mora in speech production in Japanese. Recent evidence favors the view that variance in durational measures of speech should be accounted for on a different basis across languages (Fant, Kruckenberg & Nord, 1989, 1991a,b); the rhythmic properties of Japanese are without doubt different from those of languages in which rhythm is based on stress units or syllables (Hoequist, 1983a,b).

Since English rhythm is said to be based on the stress unit, it is not surprising to find that rigid prescription of verse forms in English typically includes specification of metrical stress. Thus the limerick is defined as a verse form with the rhyme scheme *aabba* and five lines, of which the first and second each contain three stress beats, with in each case a following "silent stress" (Abercrombie, 1965), the third and fourth each have two stress beats, and the last again has three stress beats. Within this pattern, variation in number of syllables is allowed, and syllable structure is completely unconstrained. In French, on the other hand, where rhythm is said to be syllable-based, the laws of French versification can only be realized via regularity in number of syllables (de Cornulier, 1982). Thus stress units in English verse and syllables in French verse seem to play roles similar to morae in Japanese verse; versification is one way in which the basic rhythmic unit of any language is manifested.

Concepts such as phoneme, mora, syllable, and stress unit are not properties of particular languages; they are phonological constructs in terms of which any language can be described. However, just as there are differences across language communities in the way verse forms are typically constrained, and in the way phonology may be reflected in orthography, so are there differences across languages in the degree to which various phonological constructs play a role in phonological processes within the language. Likewise, there are now known to be differences across languages in the importance of the same phonological constructs in speech recognition processes.

The syllable, for instance, plays an important role in the recognition of French. In an experiment which founded an entire paradigm, Mehler, Dommergues, Frauenfelder, and Segui (1981) had French subjects listen to lists of unrelated words and press a response key as fast as possible when they heard a specified word-initial sequence of sounds. This target was either a consonant-vowel (CV) sequence such as *ba-* or a consonant–vowel–consonant (CVC) sequence such as *bal-*. The words which began with the specified target had one of two syllabic structures: the initial syllable was either open (CV), as in *balance*, or closed (CVC), as in *balcon*. Mehler *et al.* found that response time was significantly faster when the target sequence corresponded exactly to the initial syllable of the target-bearing word than when the target sequence constituted more or less than the initial syllable. Thus responses to *ba-* were faster in *balance* than in *balcon*, whereas responses to *bal-* were faster in *balcon* than in *balance*. Mehler et al. interpreted this result as evidence that French listeners segmented speech input into syllables.

Other experiments, also conducted in French, further supported this claim. Segui, Frauenfelder, and Mehler (1981) found that listeners are faster to detect syllable targets than to detect targets corresponding to the individual phonemes which make up those same syllables. Segui (1984) summarized a number of studies indicating that polysyllabic words, whether they are heard in isolation or in connected speech, are analyzed syllable by syllable. Cutler, Mehler, Norris, and Segui (1986) found that French listeners even show evidence of syllabic segmentation when listening to a foreign language (English). Thus the evidence from many studies of speech processing by French listeners suggests that their speech segmentation proceeds syllable by syllable.

However, things are different in English.

Cutler et al. (1986) found that English listeners do not show the same pattern of results in this task as French listeners. Using exactly the same experimental design as Mehler et al. (1981), but English materials (e.g., *balance, balcony*) and English-speaking subjects, they found that response time to CV (*ba-*) and CVC (*bal-*) targets was not significantly different either in *balance*- or *balcony*-type words. Nor did English listeners show evidence of syllabic segmentation when they listened to French materials (which lend themselves well to such a procedure).

The appropriate segmentation procedure for English appears to be quite different. In a stress language, such as English is, syllables can be either strong or weak; strong syllables contain full vowels, while weak syllables contain reduced vowels (usually schwa). Cutler and Norris (1988) suggested that this difference could assist segmentation. Their proposal was based on an experimental finding that listeners were slower to detect the embedded real word in, say, *mintayf* (in which the second vowel is strong) than in *mintef* (in which the second vowel is schwa). They suggested that listeners segmented *mintayf* prior to the second syllable, so that detection of *mint* therefore required combining speech material from parts of the signal which had been separated. No such difficulty would arise for the detection of *mint* in *mintef*, since the weak second syllable would not be separated from the preceding material.

Further evidence that English listeners use stress rhythm to segment speech input is found in segmentation *errors*, i.e., the way in which word boundaries tend to be misperceived. Cutler and Butterfield (1992) examined both spontaneous and experimentally elicited misperceptions, and found that erroneous insertions of a word boundary before a strong syllable (e.g., "disguise" being heard as "the skies") and deletions of a word boundary before a weak syllable (e.g., "ten to two" being heard as "twenty to") were far more common than erroneous insertions of a boundary before a weak syllable (e.g., "variability" being heard as "very ability") or deletions of a boundary before a strong syllable (e.g., "in closing" being heard as "enclosing"). This is exactly what would be expected if listeners are dealing with the segmentation problem by applying a strategy of assuming that strong syllables are likely to be word-initial, but weak syllables are not. Segmentation in English, therefore, appears to be based on the opposition of strong and weak syllables.

As Cutler, Mehler, Norris, and Segui (1992) have pointed out, the experimentally demonstrated segmentation procedures for French and English mirror each language's characteristic rhythmic structure. The use of the opposition between strong and weak syllables in segmenting English reflects the English language's characteristic stress-based rhythmic pattern, and the use of the syllable in segmenting French reflects the characteristic syllable-based rhythm of French.

What then of Japanese, in which the characteristic rhythm is mora-based? Can we expect to find that Japanese listeners make use of the mora in speech segmentation?

The purpose of the present experiments was to address these questions. By using the target-monitoring task, we extended to Japanese the series of experiments begun by Mehler et al. (1981), and continued by Cutler et al. (1986) with a comparison of English and French, and by Sebastian-Galles, Dupoux, Segui, and Mehler (1992) with Spanish and Catalan.

As we discussed above, morae in Japanese are sometimes coextensive with syllables and sometimes not. Thus the target monitoring methodology allows us to make a direct comparison between morae and syllables in the segmentation of Japanese. It should be noted that syllables may also be quite tangible constructs to Japanese language users. The *kana* orthographies are often thought of as syllabaries, and indeed, of all morae only N and Q cannot function

as separate syllables. The Chinese characters of the *kanji* orthography, moreover, when given a Chinese reading, more often represent single syllables. Thus although we have argued above that orthographic influences may render the mora a salient unit to literate Japanese, it is also true that orthographic influences may render the syllable equally salient. Thus a direct comparison between mora and syllable is of considerable interest.

As before, subjects will be asked to listen for CV and CVC targets, and the words in which these targets occur will differ in *syllable* structure: the first syllable will be either CV or CVC. (In fact, the latter structure is always CVN in the present study. N can be considered the only true syllabic coda in Japanese; Q can function as a coda but not in word-final position. Although in the type of words we used it would have been possible to use CVQ targets because the target sequence was never word-final, we chose to use one coda only.) Consider, for instance, the words *tanishi* and *tanshi*. Each has three morae: *ta-ni-shi, ta-n-shi*. But they differ in number of syllables: *ta-ni-shi* has three but *tan-shi* has two. Likewise they differ in syllabic structure; the first syllable of *tanishi* is *ta*, but the first syllable of *tanshi* is *tan*. The initial sound sequences are the same, thus satisfying the requirement that either *ta* or *tan* could serve as target for each word.

The two segmentation hypotheses propose, respectively, that the syllable and the mora play a role in speech segmentation in Japanese; and the two hypotheses predict different patterns of results in such an experiment. The syllabic hypothesis predicts the pattern of results characteristically found in the previous experiments with French listeners: CV targets (*ta*) should be detected easily in CVCVCV words (*tanishi*), and CVN targets (*tan*) should be detected easily in CVNCV words (*tanshi*), because in each case the target corresponds to the initial syllable of the stimulus word. In contrast, detection of CVN targets in CVCVCV words (*tanishi*) and of CV targets (*ta*) in CVNCV words (*tanshi*), where the target corresponds respectively to more than and less than the word-initial syllable, should be difficult. In French listeners, this difficulty expressed itself in significantly longer response times, and a similar effect in the response time pattern is one way in which Japanese listeners could exhibit the same difficulty. However, note that under a syllabic hypothesis the detection of nonsyllabic targets amounts effectively to detection of *parts of* syllables, which is assumed to occur via decomposition of initial syllabic representations into phonemes (Mehler et al., 1981). Such decomposition can, as the previous results showed, certainly be achieved by French listeners, and one reason for this could well be the availability of alphabetic orthographic representations. In Japanese, where orthographic representations are not alphabetic, it is possible that such decomposition might be hard to achieve, in which case we might predict that listeners will simply not detect a match between target and stimulus word at all, i.e., will fail to respond to nonsyllabic targets. Irrespective of how the relative difficulty expresses itself, however, it is important to note that the syllabic hypothesis predicts that CV targets in CVCVCV words and CVN targets in CVNCV words will be easy to detect, perhaps equivalently so, while CV targets in CVNCV words and CVN targets in CVCVCV words will, again, pattern similarly and will cause difficulty for listeners.

The principal prediction of the mora hypothesis, on the other hand, is that there should be no difference in responses to CV targets as a function of syllable structure; in both CVCVCV and CVNCV words, the initial CV is also the initial mora, so the CV target should have identical status in both words, and in both cases detection should present no difficulty for listeners. The predictions of this hypothesis for CVN targets are less clear-cut. Firstly, in the case of CVCVCV words, a CVN target corre-

sponds to a mora plus part of the next mora, so detection should be extremely difficult. As with the predictions from the syllabic hypothesis, the prediction from the mora hypothesis about how this difficulty will be manifested will depend on whether subjects can effectively decompose their initial moraic representations; if they can, then responses should be slow, but if they cannot, then responses should simply not occur. Secondly, in the case of CVNCV words, a CVN target corresponds to two morae. Whether or not this causes difficulty for listeners cannot be simply answered from previous work, although it is perhaps relevant that English listeners, who do not use syllabic segmentation, respond more rapidly to phoneme targets than to syllable targets (Norris & Cutler, 1988); if syllables are, to English listeners, effectively concatenations of phonemes, then this result suggests that Japanese listeners may also produce relatively slow responses to CVN targets in CVNCV words. Thus the predictions of the mora hypothesis are very different from those of the syllable hypothesis.

One problem raised by our first experiment concerned how to present the targets. In the predecessor experiments, the targets were in most cases presented visually. Thus a subject would typically see a target such as TA or TAN in upper case on a screen, immediately prior to hearing a list of words of which one might contain that target. In Japanese, however, the direct mapping of morae and syllables into the available orthographic representations means that presentation of targets either in *kana* or *kanji* would preempt the subject's decision, in that the way each input word would normally be written would fully determine whether or not a match between target and input was present.

A way around this problem, which moreover maintained comparability with the predecessor studies, was presented by the fact that all Japanese also learn Roman alphabet characters, and—especially in the case of university-educated language users such as the subjects in our experiments—are easily able to map these rapidly into sound sequences. We therefore presented the targets not in Japanese orthography at all, but in Roman characters, exactly as they had been presented in the previous experiments with European languages. Such representations should, for Japanese language users, be neutral with regard to their mapping to both morae and syllables.

In Experiment 1, therefore, we presented listeners with targets such as TA and TAN and words such as *tanishi* and *tanshi,* and measured both target detection response times, and the relative proportion of hits and misses, for each type of target in each type of word.

EXPERIMENT 1

Method

Materials

Sixteen meaningful Japanese words were chosen as stimulus words; in Roman transcription they were: *tanishi, tanshi, monaka, monka, kanoko, kanko, sanaka, sanka, nanoka, nanka, kinori, kinri, haneda, handa, shinigao, shingao.* As can be seen, the words formed eight pairs. Within each pair, the initial and final morae were identical. The second mora was either the nasal consonantal coda, or a CV mora beginning with [n]. The members of each pair were unrelated in meaning (thus *tanishi* is a kind of snail, while *tanshi* is a terminal). Japanese words may also vary in prosodic structure, to wit in placement of pitch accent (syllables bearing pitch accent are labeled H, for high, in contrast to syllables labeled L, for low). We knew of no data on the role of prosodic structure in auditory word recognition in Japanese, so we decided to incorporate prosodic structure as an additional manipulation of the materials. Thus the accent pattern of the first four pairs above is HLL, and of the last four pairs is LHH.

A further 250 words were chosen, and arranged into 64 sequences. Each target

word occurred twice, once in 1 of the first 32 sequences and once in 1 of the last 32 sequences. The sequences varied in length from 3 to 6 words, and the target words occurred in second, third, fourth, or fifth position. Of the 32 sequences which did not contain one of the experimental target words, half contained no occurrence of the specified target, while the other half contained a dummy target in varying position in the sequence. A few of the sequences without occurrence of the specified target contained a word which began with sounds similar to the target (e.g., *hoteru* with target HA, or *shimenawa* with target SHIN); such foils can prevent subjects from responding precipitately on the basis of only partial analysis of the input (Norris & Cutler, 1988).

Ten practice sequences were also constructed. These too varied in length from three to six words, and four of them had no occurrence of the specified target.

Two target orders were compiled, with type of target counterbalanced across order and first versus second half of the experiment for each item pair. Thus in order A, *tanishi* occurred with the target TA on its first occurrence and target TAN on its second, while *tanshi* occurred with target TAN on its first occurrence and target TA on its second. In order B *tanishi* occurred first with TAN and then with TA, while *tanshi* occurred first with TA and then with TAN.

The experimental and practice sequences were recorded on digital audio tape by a male native speaker of Standard Tokyo Japanese. Each sequence was preceded by its number. The words were spoken at a normal rate with approximately 2 seconds between words and approximately 10 seconds between sequences.

Subjects

Forty-eight undergraduate members of Dokkyo University took part in the experiment for a small payment. The data for 8 subjects were not used because of equipment problems. Of the remaining 40 subjects, 20 received each target order condition.

Procedure

The subjects were tested individually in a quiet room. They were instructed to listen for a word beginning with the sounds represented by the Roman characters specified as target for each sequence, and to press a response key as soon as they had detected an occurrence of this target. The target for each sequence was presented visually, on a 15 × 17-cm card, immediately prior to the beginning of the sequence.

The sequences were presented over headphones from a DAT recorder. The output from this recorder was also fed via a mixer to a second DAT recorder which via the same mixer also recorded a pulse triggered by the subject's response.

The intervals between onset of the target word and response pulse were measured individually for each subject on a Kay SonaGraph 5500 to ascertain reaction times.

Results

After the experiment had been run it was discovered that the target assignment for one item had been erroneous, resulting in no data being available for one of the conditions for that item. Therefore it was decided to omit that item and its pair from the analysis.

Mean number of missed responses and mean response times (RTs) were determined for each subject and each item, and separate analyses of variance were conducted on each measure with subjects and with items as random factors.

The most immediately striking feature of the results was apparent in the analysis of missed responses. The results of this analysis are shown in Fig. 1. It can be seen that in three conditions subjects missed less than 8% of targets, but in the fourth condition—CVN targets with CVCVCV words—the miss rate was extremely high: 64.3%.

Analyses of variance showed significant

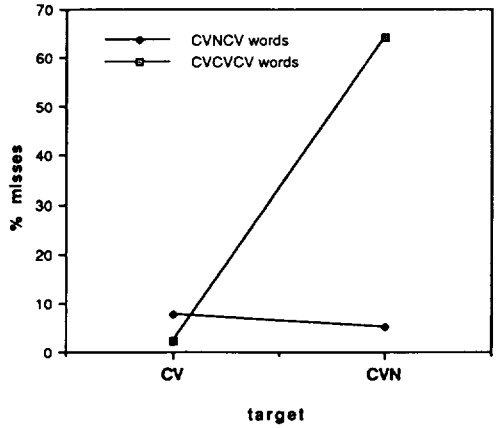

FIG. 1. The mean number of missed targets as a function of the size of a visual target sequence (CV, e.g., *ta-* versus CVC, e.g., *tan-*) and the phonological structure of a stimulus word (CVCVCV, e.g., *tanishi* versus CVCCV, e.g., *tanshi*); Experiment 1.

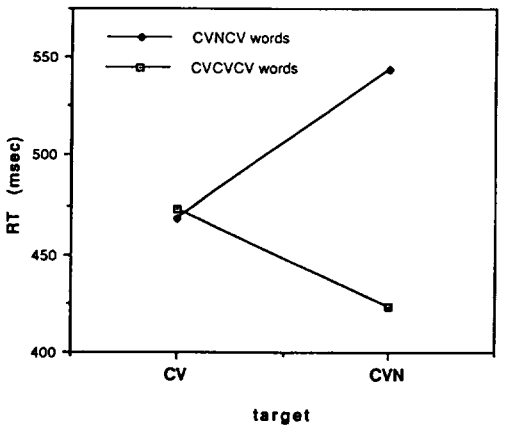

FIG. 2. The mean target detection response time (RT) in ms as a function of the size of a visual target sequence (CV, e.g., *ta-* versus CVC, e.g., *tan-*) and the phonological structure of a stimulus word (CVCVCV, e.g., *tanishi* versus CVCCV, e.g., *tanshi*); Experiment 1.

effects in both analyses for the word type comparison ($F1$ [1,38] = 158.23, $p < .001$; $F2$ [1,10] = 77.23, $p < .001$), the target size comparison ($F1$ [1,38] = 166.53, $p < .001$; $F2$ [1,10] = 93.82, $p < .001$), and the interaction of these two factors ($F1$ [1,38] = 147.09, $p < .001$; $F2$ [1,10] = 112.47, $p < .001$). To examine the basis for the interaction, we conducted t tests for the two word types separately; these showed that in CVCVCV words the miss rate for CVN targets was significantly higher than it was for CV targets ($t1$ [39] = 14.54, $p < .001$; $t2$ [6] = 11.34, $p < .001$), while in CVNCV words there was no significant difference between target types.

The mean RTs are shown in Fig. 2. Analyses of variance showed that the only RT effects which were significant in both subjects and items analyses were the main effect of word type, with responses to CVCVCV words being faster than responses to CVNCV words ($F1$ [1,38] = 16.65, $p < .001$; $F2$ [1,10] = 10.62, $p < .001$), and the interaction between word type and target size ($F1$ [1,38] = 25.67, $p < .001$; $F2$ [1,10] = 42.22, $p < .001$). However, the validity of an overall analysis is very dubious when one condition contains markedly fewer responses than the other three conditions. Omission of the CVN target/CVCVCV word condition, and analysis via t tests of the relationship between the remaining three conditions showed that RTs to CV targets were not significantly different in the two word types; but in CVNCV words RTs to CVN targets were longer than RTs to CV targets ($t1$ [39] = 6.01, $p < .001$; $t2$ [6] = 13.84, $p < .001$).

Although a comparison between HLL and LHH accent patterns was included in both miss-rate and RT analyses, it was responsible for no significant effects, either alone or in interaction.

Discussion

The results from both analyses provide evidence which bears on the hypotheses proposed above. The predictions of the syllable hypothesis for Japanese are clearly disconfirmed. This hypothesis predicted that CV targets in CVCVCV words and CVN targets in CVNCV words should both be easy to detect, and equivalently so, while CV targets in CVNCV words and CVN targets in CVCVCV words should both be hard to detect, and equivalently so. In fact, CV targets in CVNCV words were as easy to detect as CV targets in CVCVCV

words, as shown both by the miss-rate analysis and the RT analysis, while in the RT analysis CVN targets in CVNCV words in fact produced the slowest responses. Thus the syllabic hypothesis cannot account for Japanese listeners' performance.

The mora hypothesis, on the other hand, can account for the results. Because the initial mora of both word types is the same (CV), the mora hypothesis predicted that CV targets in both CVCVCV and CVNCV words should be easy to detect, and equivalently so. Indeed, there is no difference between these two conditions either in miss rates or in RTs. This hypothesis predicted that CVN targets should be hard to detect in CVCVCV words, and the hit/miss analysis showed that they were; it also predicted that CVN targets in CVNCV words, which were effectively a simultaneous match to two morae, would (assuming the generalizability of the results of Norris & Cutler (1988) from English) produce responses, but these would be relatively slow, and the analyses showed that this too was the case—the miss rate in this condition was low, but RTs were high.

One apparently anomalous feature of the results is the RTs which were produced to CVN targets in CVCVCV words. Both hypotheses predicted that responses in this condition, if made, should be slow; as Fig. 2 shows, they were not. Statistically, RTs in this condition did not differ from RTs to CV targets in the same words. However, the first thing to note is that because of the high miss rate in this condition these responses are in fact very few: only 36% of possible responses in this condition. They are fast simply because they tend to be produced by fast subjects only. In an analysis in which the subjects were divided into quartile subgroups based on their mean RT, the pattern of missed responses differed across subgroups: specifically, the number of missed CVN targets in CVCVCV words was greatest for the slowest quartile and was least for the fastest quartile. (However, the pattern of RT results was exactly the same over all the subgroups; from the fastest to the slowest quartile, it was always the case that reaction times were relatively slow to CVN targets in CVNCV words, and essentially the same in the other three conditions. That is, the fastest and slowest subjects differed only in the proportion of responses contributed per condition, not in the way the detection task was being performed.)

Our interpretation of the relatively few responses to CVN targets in CVCVCV words is that they are in fact simply false alarms; subjects are responding on the basis of a match of the first part of the target alone. One reason may be that the subject's representation of the target is, mistakenly, CV instead of CVN (either because the target was misperceived or because it has become corrupted in memory); in this case it would not be surprising that the responses in this condition are virtually identical to the responses in the CV target condition with the same words. We conducted further analyses in which we compared just those responses which occurred in the CVN target condition with CVCVCV words with their pair trials (same subject, same item, but with CV target). The overall mean RTs for these two sets were within 4 ms of one another whether calculated across items or across subjects, and there was no consistent pattern to the source of these responses across items. Thus it would seem highly likely that responses to CVN targets in CVCVCV words are simply isolated errors made by subjects trying to respond as fast as they can, and perhaps responding mistakenly to only the CV portion of the target.

The pattern of results in this experiment thus appears to offer strong support for the mora hypothesis but none to the syllable hypothesis. The possibility remains, however, that the results in some way reflect acoustic properties of the materials. It may be the case, for instance, that CVN targets are responded to more slowly in CVNCV words because the CVN portion of these

words is particularly long; likewise it may be the case that CVN targets are missed more often than not in CVCVCV words because the nasal consonant in these words is unclearly articulated or in some other way hard to identify as such. A simple way to test these possibilities is to present the same materials to nonnative listeners. If such acoustic factors have influenced the results of Experiment 1, they should exercise similar effects on the performance of other listeners regardless of native language. In Experiment 2, therefore, we presented the same materials to listeners whose native language was not Japanese but English.

EXPERIMENT 2

Method

Materials

The materials were as in Experiment 1. However, the tape was copied and in the process the numerals at the start of each sequence were removed, since these would essentially be nonwords to English listeners and hence would merely function to lengthen each sequence by one item. A timing mark was placed on the second channel of the tape aligned roughly with the onset of each target item.

Subjects

Twenty-four members of the Applied Psychology Unit subject panel, ranging in age from 19 to 35 years, took part in the experiment for a small payment. None had any knowledge of Japanese. Twelve were assigned to each target order condition.

Procedure

The subjects were tested individually in a sound-dampened room. They were presented with instructions which described the task and illustrated the materials with Japanese words and names which would be known to English subjects (e.g., "you might see the target SON and hear *tempura Mitsubishi sukiyaki Sony*"). One deliberately wrong example ("FU as in *futon*") was included in the instructions as a potential diagnostic—although the sound sequence represented by FU occurs in the English pronunciation of *futon*, it does not actually occur in the Japanese pronunciation, and any subject who pointed that out would be revealed as knowledgeable about Japanese (in fact, no subject made such a comment).

The sequences were presented over headphones from a DAT recorder. The targets were displayed in upper case on a VDU screen. Target presentation, timing and response collection were under the control of a Zenith microcomputer running the TSCOP experimental software (Norris, 1984).

The intervals between target onset and timing mark were measured and the response times adjusted by these amounts to give responses from exact target onset.

Results

Mean number of missed responses and mean RTs were determined for each subject and each item, and separate analyses of variance were conducted with subjects and with items as random factors. The mean number of missed targets for each condition is shown in Fig. 3.

It can immediately be seen that the striking asymmetry evident in Figure 1 is not replicated here. In fact the English subjects were extremely good at correctly detecting the targets in these Japanese words: the miss rate was between 5 and 9% in all four conditions. The analyses of variance conducted on the number of missed targets showed no main effects or interactions significant in either analysis.

The mean RTs for each condition are shown in Fig. 4, and again it can be seen that the response pattern is quite different from that shown by the Japanese listeners in Experiment 1. The analyses of variance conducted on the RTs showed no main effects or interactions significant in both anal-

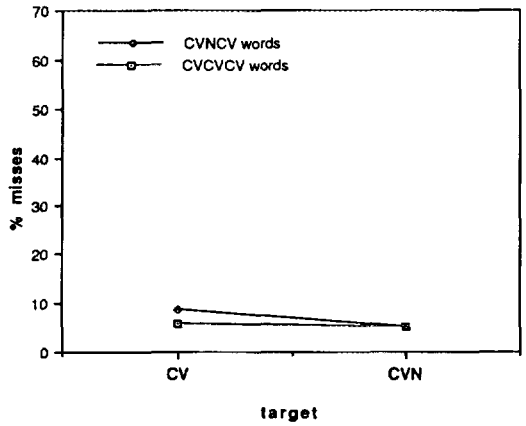

FIG. 3. The mean number of missed targets as a function of the size of a visual target sequence (CV, e.g., *ta-* versus CVC, e.g., *tan-*) and the phonological structure of a stimulus word (CVCVCV, e.g., *tanishi* versus CVCCV, e.g., *tanshi*); Experiment 2.

yses; specifically, in the analysis by subjects none of the three effects (word type, target size, prosodic structure) reached significance either alone or in interaction, while in the analysis by items the target size effect (faster RTs to CV than to CVN targets) was significant ($F2$ [1,28] = 5.14, $p < .04$) but no other main effect or interaction was significant.

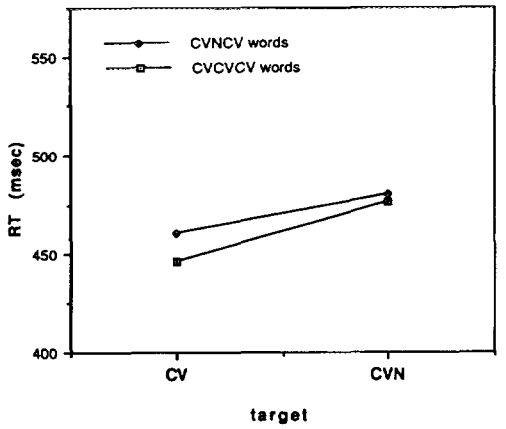

FIG. 4. The mean target detection response time (RT) in ms as a function of the size of a visual target sequence (CV, e.g., *ta-* versus CVC, e.g., *tan-*) and the phonological structure of a stimulus word (CVCVCV, e.g., *tanishi* versus CVCCV, e.g., *tanshi*); Experiment 2.

Discussion

These results do not at all resemble the results produced by the native listener group. Most importantly, miss rates were not particularly high for CVN targets in CVCVCV words (or for any other condition, for that matter). Likewise, the RTs did not pattern like the Japanese listeners' RTs. From this we may conclude that the results of Experiment 1 were not due to some acoustic properties of the materials which would exercise the same effect on any listener group independent of native language. Instead, the earlier results may properly be ascribed to characteristics of native language processing procedures. We argued above that the mora hypothesis best fits the results of Experiment 1; Japanese listeners use mora structure in speech segmentation. We may now conclude that English listeners do not exploit mora structure in the same way.

Note that the syllable hypothesis also does not predict the results for English listeners. This is in agreement with previous studies in which English listeners have been shown not to use syllabic segmentation when listening either to their native language or to French (Cutler et al., 1986). In previous experiments, in contrast, English subjects have shown a consistent tendency to respond faster (to any target) in words beginning CVCV than in words beginning CVCC (Cutler et al., 1986; Cutler, Norris & Williams, 1987); though the same tendency was weakly apparent here, it was not statistically significant.

Thus Experiment 2 has ruled out one artifactual account of the results of Experiment 1. It may still be objected, however, that an element of unnaturalness was introduced into Experiment 1 by the method of target presentation. Although all Japanese students read Roman characters easily, and writing Japanese words (e.g., names) in Roman characters is quite common, this is of course not the orthography which is normally used for Japanese text. It is certainly

conceivable that the task of constructing a target representation from Roman characters to be matched against spoken Japanese words could have introduced into the experiment some artifact which in turn could have influenced the results; in particular, it is conceivable that the orthographic representation may have been converted by the subjects into a mora-based (*kana*) representation of the targets. Thus it is important to test whether the results of Experiment 1 will generalize to alternative modes of target presentation. In Experiment 3, therefore, we replicated Experiment 1 exactly except in target presentation mode: the targets were presented auditorily.

EXPERIMENT 3

Method

Materials

The materials were the same as for Experiment 1. The recording used for Experiment 1 was copied and auditory target specifications were added prior to each sequence. These target specifications were spoken separately rather than edited from the experimental words (it has been shown that such sequences extracted from Japanese speech are very difficult to recognize; Kuwabara, 1982).

Subjects

Forty-three undergraduate members of Dokkyo University took part in the experiment for a small payment. The data for 3 subjects were lost due to equipment failure. Of the remaining 40 subjects, 20 received each target order.

Procedure

The subjects were instructed to listen for a word beginning with the sounds specified as target for each sequence, and to press a response key as soon as they had detected an occurrence of this target. The subjects were tested in pairs, and the materials were presented from an analogue tape recorder. The procedure was in other respects the same as for Experiment 1.

Results

The mean number of missed responses and the mean RTs were determined for each subject and each item, and separate analyses of variance were conducted on each measure with subjects and with items as random factors. Figure 5 shows the mean number of missed responses. It can be seen that the distinctive asymmetry evident in Fig. 1 is replicated in this second experiment with Japanese listeners. In three conditions the miss rate was 10% or less; but CVN targets in CVCVCV words were missed on 64.5% of trials.

Analyses of variance showed the same pattern of results as in Experiment 1: main effects for word type ($F1$ [1,38] = 151.48, $p < .001$; $F2$ [1,10] = 53.82, $p < .001$), target size ($F1$ [1,38] = 148.68, $p < .001$; $F2$ [1,10] = 44.5, $p < .001$) and the interaction of these two factors ($F1$ [1,38] = 181.09, $p < .001$; $F2$ [1,10] = 60.67, $p < .001$). T tests again showed no difference between the two target types in CVNCV words, but in CVCVCV words CVN targets were missed significantly more often than CV targets ($t1$ [39] = 15.12, $p < .001$; $t2$ [6] = 8.48, $p < .001$).

The mean RTs across conditions are shown in Fig. 6; again, the RT analysis

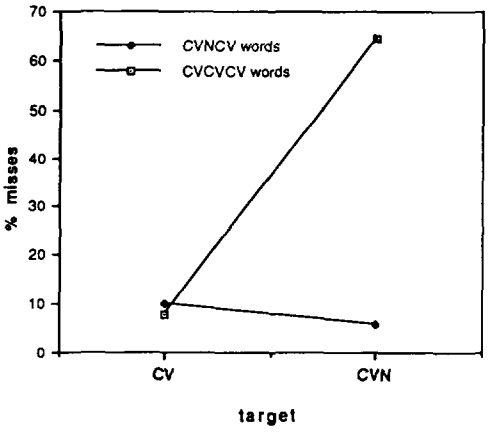

FIG. 5. The mean number of missed targets as a function of the size of an auditory target sequence (CV, e.g., *ta-* versus CVC, e.g., *tan-*) and the phonological structure of a stimulus word (CVCVCV, e.g., *tanishi* versus CVCCV, e.g., *tanshi*); Experiment 3.

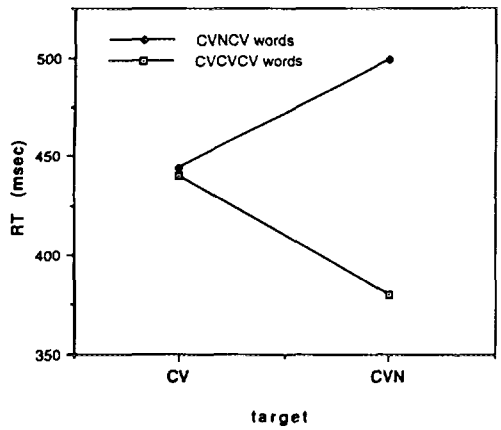

FIG. 6. The mean target detection response time (RT) in ms as a function of the size of an auditory target sequence (CV, e.g., *ta-* versus CVC, e.g., *tan-*) and phonological structure of stimulus word (CVCVCV, e.g., *tanishi* versus CVCCV, e.g., *tanshi*); Experiment 3.

showed the same pattern as in Experiment 1. The only two variables to reach significance in both subjects and items analyses were the effect of word type, with CVCVCV words being responded to faster than CVNCV words ($F1$ [1,38] = 22.33, $p < .001$; $F2$ [1,10] = 14.44, $p < .005$), and the interaction between word type and target size ($F1$ [1,38] = 29.32, $p < .001$; $F2$ [1,10] = 32.23, $p < .001$). Again we conducted t tests on the three conditions in which few targets were missed, omitting the CVN target/CVCVCV word condition. These analyses showed no difference between responses to CV targets as a function of word type; but in CVNCV words CVN targets were responded to more slowly than CV targets ($t1$ [39] = 4.47, $p < .001$; $t2$ [6] = 2.78, $p < .04$).

As in Experiment 1, a further analysis was conducted in which the subjects were divided, on the basis of their overall mean RT, into subgroups; once again, there was no difference between the subgroups in the pattern of RTs. The pattern of missed responses, however, did differ across subgroups, and in exactly the same manner as we observed in Experiment 1: the responses to CVN targets in CVCVCV words were contributed to the greatest extent by those subjects with the fastest mean RTs. Once again, therefore, we would argue that the few responses which occur in this condition are in fact false alarms.

To check that the results of Experiments 1 and 3 were indeed directly comparable, we conducted a joint analysis of the two experiments. The additional variable of Experiment (visual versus auditory targets) was not itself significant, and entered into no significant interactions, either in the joint analysis of miss rates or of RTs.

Discussion

The replication of Experiment 1's results strongly confirms our conclusions from the preceding experiments: Japanese listeners do not naturally segment speech syllable by syllable; they do naturally segment it mora by mora. The change in mode of target specification made no difference to the pattern of results, either in the hit/miss analysis or the response time analysis.

Experiment 2 has already established that no artifact of the materials underlies the results—the responses of English listeners to these words in no way resembled the responses of native listeners. However, we decided to conduct a further replication experiment with nonnative listeners. In Experiment 4 we presented the same materials to French listeners. Recall that all previous experiments with French listeners (Mehler et al., 1981; Cutler et al., 1986) had shown a distinctive pattern associated with a syllabic segmentation of the speech input: CV targets were easier to detect when the initial syllable of the target item was CV rather than CVC, but CVC targets were easier to detect when the target item's initial syllable was CVC rather than CV.

The opportunity for precisely this pattern of results also exists with the materials of Experiments 1 to 3. Syllabic segmentation would result in the target *ta-* being easier to detect in *tanishi* than in *tanshi,* while the target *tan-* would be easier to detect in *tanshi* than in *tanishi.* Cutler et al. (1986) showed that French listeners produce this characteristic pattern of results with for-

eign-language (in that case, English) speech input; in Experiment 4 we tested whether French listeners would also produce this response pattern with Japanese.

EXPERIMENT 4

Method

Materials

The materials were as in Experiment 3, except that three filler sequences were omitted. The sequences were digitized at 16 kHz via an OROS AU 22 A/D converter, and stored in a Toshiba T5200 microcomputer. The numerals at the start of each sequence were removed for the same reason as in Experiment 2. Using the speech editor MANITOU, a timing mark was aligned with the onset of each target item.

Subjects

Thirty-three native French-speaking volunteer subjects (students and members of the University of Paris community, ages between 20 and 30), took part in the experiment. Only one target order was used.

Procedure

The subjects were tested individually in a sound-dampened room. They were told that they would hear short sequences of Japanese words preceded by two occurrences of a target specification, and they were instructed to listen for a word beginning with the sounds specified as target for each sequence, and to press a response key as soon as they had detected an occurrence of this target.

The targets and sequences were presented over headphones. Each trial began with a 200-ms auditory warning, followed by two presentations of the target, and then by the sequence of words, with an interword interval of 700 ms. Audio presentation, timing, and response collection were under the control of the Toshiba microcomputer. For eight subjects on four trials the target items were accidentally preceded by a pause in the computer presentation; these trials were omitted from the analysis for these subjects.

Results

Mean number of missed responses and mean RTs were determined for each subject and each item, and separate analyses of variance were conducted with subjects and with items as random factors.

The mean number of missed targets for each condition can be seen in Fig. 7. As in Experiment 2, the characteristic pattern shown by Japanese listeners is not replicated with nonnative listeners (in fact, once again, the nonnative listeners had an overall lower miss rate than the native listeners of Experiments 1 and 3). Instead, the pattern we see is the distinctive crossover interaction which is predicted with syllabic segmentation. In the two conditions in which the syllabic hypothesis predicts that target detection will be easy, the miss rate is around 10% (i.e., around the same level as for English listeners in all four conditions and for Japanese listeners in three), while in the two conditions in which the syllabic hypothesis predicts that target detection will be hard, the miss rate is between 20 and 30%. Analyses of variance showed that the

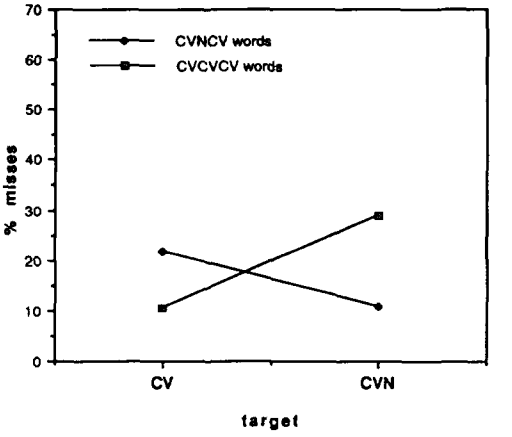

FIG. 7. The mean number of missed targets as a function of the size of an auditory target sequence (CV, e.g., *ta-* versus CVC, e.g., *tan-*) and the phonological structure of a stimulus word (CVCVCV, e.g., *tanishi* versus CVCCV, e.g., *tanshi*); Experiment 4.

crossover interaction was significant in both analyses ($F1$ [1,32] = 29.25, $p < .001$; $F2$ [1,14] = 26.1, $p < .001$); neither main effect was significant. T tests on the components of the interaction revealed that in CVCVCV words the miss rate for CVN targets was significantly higher than for CV targets ($t1$ [32] = 4.1, $p < .001$; $t2$ [7] = 4.79, $p < .002$), while in CVNCV words the miss rate for CV targets was significantly higher than the miss rate for CVN targets ($t1$ [32] = 3.79, $p < .001$, $t2$ [7] = 2.56, $p < .04$).

The mean RTs for each condition are shown in Fig. 8; again the pattern is unlike that shown by native listeners. As predicted by the syllabic hypothesis, RTs in CVCVCV words are faster for CV targets than for CVN targets, while RTs in CVNCV words are faster for CVN than for CV targets; the crossover pattern is not evident, however, because there is in this case an additional effect of word type: CVCVCV words are responded to faster than CVNCV words. The analyses of variance conducted on the RTs showed that this main effect of word type was the only effect (or interaction) to approach our criterion of significance ($F1$ [1,32] = 8.75, $p < .01$; $F2$ [1,14] = 4.17, $p < .06$). However, although the interaction between word type and target type was not significant ($F1$ [1,32] = 3.41, $p < .08$; $F2$ [1,14] = 1.06), CV targets were responded to significantly faster in CVCVCV words than in CVNCV words ($t1$ [32] = 3.17, $p < .005$; $t2$ [7] = 2.13, $p < .075$). This effect is predicted by the syllabic hypothesis, but runs counter to the predictions of the mora hypothesis; it was not shown by the Japanese subjects in either Experiment 1 or Experiment 3, or by the English subjects of Experiment 2.

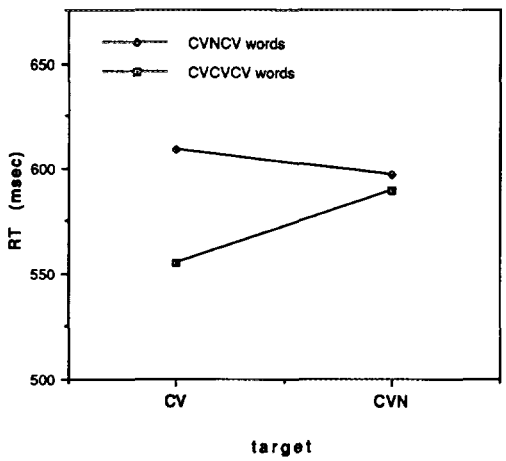

FIG. 8. The mean target detection response time (RT) in ms as a function of the size of an auditory target sequence (CV, e.g., *ta-* versus CVC, e.g., *tan-*) and the phonological structure of a stimulus word (CVCVCV, e.g., *tanishi* versus CVCCV, e.g., *tanshi*); Experiment 4.

Discussion

The results of Experiment 4 show that the response patterns of French listeners are, as predicted, best accounted for by the syllabic hypothesis; CV targets are responded to more accurately in CVCVCV words than in CVNCV words, while CVN targets are responded to more accurately in CVNCV words than in CVCVCV words. Although the RT analysis did not in this case produce a significant crossover interaction, the predictions of the syllable hypothesis were even here partially confirmed. Thus the results of Experiment 4 present further confirmation of Cutler et al.'s (1986) result: even with foreign-language input, French listeners will attempt to apply syllabic segmentation where the opportunity for it seems to exist. Syllabic segmentation thus appears to be the natural pattern which French listeners apply to speech input, whether in their own language, in a partially familiar language (English), or in an almost wholly unfamiliar language (Japanese).

The results also further confirm our conclusions from Experiments 1 and 2. Nonnative listeners did not replicate the pattern of results shown with the same materials by native Japanese listeners. Thus the native pattern can not be ascribed to some acoustic or other low-level characteristic of the stimulus materials. Instead, it must reflect aspects of native speech recognition proce-

dures; as we have argued, it is best interpreted as reflecting segmentation procedures based on the mora.

General Discussion

In four experiments on the detection of sub-word targets in Japanese words we have shown that the patterns of performance produced by native and by nonnative listeners are fundamentally different. These experiments thus add to the already substantial body of evidence that response patterns in speech target detection experiments are language-specific. Previous research has shown that closely matched experimental conditions produce quite different patterns of responding by English and French listeners (Cutler et al., 1986), and by Spanish and Catalan listeners (Sebastian et al., 1992). The present study has shown that Japanese listeners produce yet another pattern of response, which differs from that of both English and French listeners.

This series of experiments has as its overall aim elucidation of the processes which intervene between perception of an acoustic speech input and contact with stored representations of words. It is assumed that the many factors which produce variability in speech rule out direct mapping of acoustic input onto the lexicon; instead, the input must be transformed into an abstract prelexical representation. Not only must this representation be abstract, but it must be discrete; that is, the input must be segmented into discrete units at this prelexical representation stage. It is the nature of the units of segmentation which is at issue.

In many ways the syllable seems like a good candidate for a universal unit of representation, and initial studies in this domain suggested that it indeed played this role (Mehler et al., 1981; Segui et al., 1981). The syllable, after all, is a unit in terms of which the phonology of any language can be described; and moreover, those initial studies produced results highly consistent with syllable-sized prelexical units of representation. However, these studies were all conducted in French; later research destroyed the syllable's claim to universality in prelexical representations by showing results inconsistent with this claim for English (Cutler et al., 1986). Now the present study has produced further evidence inconsistent with syllabic prelexical representations, this time in Japanese. The syllable hypothesis did not correctly predict the response pattern of Japanese listeners. First, the response pattern to CV targets, both in terms of RTs and number of missed responses, was essentially identical in CVCVCV and CVNCV words, although the syllable hypothesis would predict a disadvantage for CV targets in CVNCV words. Second, the reaction times to the two target types in CVN words patterned in exactly the opposite direction to that predicted by the syllable hypothesis.

We argued that the observed pattern of responses by Japanese listeners with the materials of our experiments can, instead, best be described in terms of a mora-sized unit of segmentation in this language. The response pattern to CV targets, both in miss rates and RTs, is essentially identical in both types of word; this is exactly as predicted by the mora hypothesis, because in both CVCVCV and CVNCV words the initial mora is CV. Likewise, the mora hypothesis can account for why the response pattern to CVN targets differs across the two word types. In CVNCV words, CVN targets are responded to, but responses are long; this is as would be predicted if CVN amounts to a complex (i.e., two-mora) target. CVN targets in CVCVCV words, however, are more often than not simply missed; again, this is as predicted since at the mora level the target and stimulus do not match. The fact that response difficulty (i.e., the case of CVN targets in CVCVCV words) is manifested in Japanese listeners by missed responses rather than by long RTs suggests, indeed, that the Japanese listeners simply did not decompose morae at all; they segmented the spoken words mora

by mora, they may also have represented the targets as complete morae, and their response patterns were determined by mora structure.

Two qualifications to this conclusion may be warranted. Firstly, we note that the only mora structures in our experimental words were CV (admittedly the most common mora structure) and the nasal coda N. Although our results are certainly suggestive of mora-based responding, we cannot as yet be certain that the same pattern will be found with the other three mora structures (V, CCV, and Q); nor can we be certain that the pattern we have found with the two-mora syllable structure CVN will generalize to the other two-mora syllable structures which Japanese allows. Further experiments are necessary to resolve this issue.

Second, it is appropriate to consider whether the close relationship between Japanese phonology and (one type of) orthography may have played a role in our findings. The mora is properly a phonological construct, and our results therefore suggest that Japanese listeners are using the phonological structure of their language in a way very similar to the phonological procedures which our previous studies have suggested for French and English listeners. As we described in the Introduction, however, the mora is also an orthographically real construct to our Japanese subjects, in that the *kana* orthographies directly encode mora structure. The possibility exists, therefore, that our subjects may have produced their moraic response patterns via an orthographic representation rather than a phonological representation. This would require first that subjects formed an orthographic representation of the specified target, in terms of the mora-based *kana* orthography, and second that this representation could be directly matched against an orthographic representation of each incoming word, again in *kana*. As to the first step, target representation, we have no evidence as to whether subjects did this, but it is presumably possible. Both the target specifications of Experiment 1 (Roman alphabetic characters) and of Experiment 3 (spoken) could have allowed a translation into *kana* representations, either as sole representation or in conjunction with a phonological representation. The second step, *kana* encoding of the spoken words, is rather more contentious. The *kana* orthographies are used in Japanese text for representation of (a) closed class words and inflectional affixes (in *hiragana*), and (b) foreign words (in *katakana*). Our materials contained no closed class words, no inflected words, and no foreign words. All of the words (experimental or filler) in our lists would normally be written with a *kanji* representation, and if Japanese listeners are asked to visualize such a word in writing, it is the *kanji* which they visualize (indeed, Japanese speakers often sketch *kanji* characters in the air to disambiguate homophones in conversation).

For words which are normally written as *kanji*, construction of a *kana* representation implies conscious representation of the phonological pattern of the word. Thus one might argue that orthographically based responding seems to involve construction of a secondary representation for performing the task via a primary representation (i.e., the conscious phonological representation) which itself could have served as the basis for a response; such a claim would seem to require some independent motivation, and we do not know of any. Nevertheless, the mediation of a *kana* orthographic representation in our subjects' responses is difficult to rule out. For instance, it is not feasible to conduct an experiment to test for the presence of orthographic involvement in detection responses in the way in which one might in a language such as English, with an orthography in which graphemes encode (more or less well) individual phonemes. Here one would ask whether listeners were using an orthographic code in phoneme detection by looking at cases in which the mapping in one direction or other was im-

perfect. For instance, one could ask whether responses would be slower to noncanonical representations of a particular phoneme—e.g. to /s/ in *century* as opposed to *settler;* or one could ask whether false alarm responses might occur to graphemes with no corresponding phonological representation—e.g. to /k/ in *knee* or to /p/ in *psychology*. Because the mapping of phonology to the *kana* orthographies is virtually exact, however (see footnote 1 for a discussion of the very few exceptions), it is in practical terms impossible to ask either type of question of morae in Japanese.

One possible test which presented itself was to investigate the role of word familiarity; this plays a strong role in *kanji* recognition, and it may as a result be effectively easier for subjects to construct *kana* representations of unfamiliar words than of familiar words. The materials used in these experiments covered a fairly large range of frequency of occurrence, so we were able to compare the pattern of responding to more familiar versus less familiar words. We found no difference at all as a function of familiarity.

Whether the code which Japanese listeners used in Experiments 1 and 3 was phonological, orthographic, or a bit of both, the important finding is that, with CV and N structures at least, a mora-based code seems to be easy and natural for listeners to use. The principal argument which we would raise against a *purely* orthographic account of our results is that it presents no parallel with the existing results from other languages. A phonological explanation, on the other hand, suggests a clear parallel between the mora in Japanese, the syllable in French, and stress units in English. As we discussed in the introduction, in Japanese the mora is the unit of *rhythm*. In French the unit of rhythm is the syllable; in English the unit of rhythm is the stress unit. We also described in the introduction how our previous results indicated that speech segmentation can be based on the syllable in French and on the stress unit in English. The present results suggest that speech segmentation can be based on the mora in Japanese. In other words, this would constitute three separate cases in which under appropriate conditions speakers of a particular language can base speech segmentation upon units which happen also to be the characteristic units of rhythm for their language.

We reiterate that all of these units are abstract linguistic constructs which in principle can be used to describe any language. Of course, the utility of each construct varies widely from language to language. To take a gross example, consider the concept of tone; this is clearly meaningful when applied to languages like Chinese, Thai, and Igbo which have lexical tone. It is also meaningful to ask whether, for instance, English, French, and Japanese have tone; but having learned that the answer is no, the researcher would not gain any advantage by trying to use the construct in describing any of these languages. The situation with rhythmic constructs is exactly analogous. It makes sense to ask of any language, for instance, whether it has stress contrasts; if the answer is yes (as it is for English), then there is sense in describing the stress units, but if the answer is no (as it is with French and Japanese), then there is little point in applying the relevant constructs. Thus, for instance, while it may be true to say that all syllables in French and Japanese speech are normally strong, it is a redundant exercise, because the concept is irrelevant in the description of these languages.

Likewise, we need to ask about each such concept not only whether it is relevant in the proper phonological description of a language's rhythmic structure, but also, and separately, whether it plays a role in speech segmentation for the community of speakers of that language. It is, therefore, striking to find three cases in which the answer, for a given unit in a given language, is yes on both counts. This suggests a clear pattern: the way listeners accomplish

speech segmentation is influenced by the way their native language organizes its rhythm.

Why should this be so? We suggest that the answer may well be found in the earliest stages of language acquisition. Consider the infant's task in learning to distinguish meaning in the speech signals which occur in its environment. The segmentation problem which we described above—the necessity of isolating discrete units in the continuous speech signal—is compounded; the infant has no existing store of meaningful units. The infant's task is, indeed, to *build* a lexicon. On what basis can this process be started?

To answer this question we need to know what characteristics of speech signals are particularly salient to infant perception. Unsurprisingly, some components of linguistic rhythm prove to be salient. Infants as young as 4 days old are sensitive to the number of syllables in an utterance (Bijeljac-Babic, Bertoncini & Mehler, in press), in that they can discriminate a sequence of three-syllable words from a sequence of two-syllable words, although they cannot at the same age discriminate between sequences differing in number of phonemes. Infants of two months can distinguish bisyllables with initial stress from bisyllables with final stress (Jusczyk & Thompson, 1978; Spring & Dale, 1977), although under at least some conditions they fail to distinguish phonemic contrasts within such sequences (Karzon, 1985); moreover, the stress change discrimination appears to be far more robust than those phonetic discriminations of which infants at that age are capable (Jusczyk & Thompson, 1978).

The importance of the lexicon-building task (and, of course, the relative speed and ease with which it is accomplished by human infants) renders it likely that the newborn child is equipped with procedures which enable it to focus efficiently upon aspects of speech which assist in segmenting the continuous speech stream into meaningful units. The characteristic rhythmic pattern of a language is certainly one such aspect. The child's endowment in this respect might be thought of as a process of determining the smallest occurring level of regularity in the speech input with which it is presented. Perceptual processing procedures could then be refined to exploit as efficiently as possible the regularity at that particular level. Of course, any human infant can acquire any human language. Thus it is not surprising to find that a variety of rhythmic regularities can be discriminated by the newborn infant. But once the characteristic rhythmic structure of the speech input has been identified, this structure can form the basis of segmentation procedures by which potential meaningful units in the continuous speech stream can be identified—in other words, it can provide the framework for the lexicon-building process. *Prosodic bootstrapping* is the term which has been applied to the infant's reliance on prosodic structure in such tasks.

We would expect, then, that as language acquisition progresses, the young child should lose its newborn impartiality toward the range of possible rhythmic structures available in language, and begin to exhibit preferential behavior. Indeed, even at 9 months of age, children acquiring English show a preference for more typical stress patterns of English words over less typical (Jusczyk, Cutler, & Redanz, 1993), which suggests at least that stress is a salient attribute in the lexical stock which children at this age are engaged in assembling. Certainly such preferences exist by the time a child has become a relatively competent language user, i.e., by 3 years of age and more, and they form the basis for explicit segmentation choices. Thus young French-speakers are able to use syllables in segmentation (Alegria, Pignot, & Morais, 1982), young English-speakers are able to use stress (Peters, 1985; Gerken, 1993), and young Japanese-speakers are able to use morae (Mann, 1986).

We suggest, however, that only one type of procedure will be developed by which

rhythmic structure is exploited to aid speech segmentation. Our previous work has shown that, for instance, English-speaking monolinguals cannot exploit syllabic rhythm to aid segmentation, even when they are listening to speech in French, in which syllabic segmentation would be efficient (Cutler et al., 1986). More remarkably yet, maximally competent French–English bilinguals appear to have only one such segmentation procedure available to them (Cutler et al., 1992)—either the syllabic segmentation procedure typical of French monolinguals, or the stress-based segmentation procedure typical of English monolinguals. Such segmentation procedures may be an integral part of a speaker's linguistic competence, just as a language's characteristic rhythm is a fundamental aspect of its phonological structure; at the earliest stages of language acquisition they are crucially involved in enabling the construction of a lexicon, and in later life they continue to serve as useful aids in the process of speech segmentation.

REFERENCES

ABERCROMBIE, D. (1965). Syllable quantity and enclitics in English. In *Studies in Phonetics and Linguistics*. Oxford: Oxford Univ. Press.

ALEGRIA, J., PIGNOT, E., & MORAIS, J. (1982). Phonetic analysis of speech and memory codes in beginning readers. *Memory & Cognition*, **10**, 451–456.

BECKMAN, M. E. (1982). Segmental duration and the "mora" in Japanese. *Phonetica*, **39**, 113–135.

BERTINETTO, P. M., & FOWLER, C. A. (1989). On sensitivity to durational modifications in Italian and English. *Rivista di Linguistica*, **1**, 69–94.

BIJELJAC-BABIC, R., BERTONCINI, J., & MEHLER, J. (1992). How do 4-day-old infants categorize multisyllabic utterances? In press, *Developmental Psychology*.

DE CORNULIER, B. (1982). *Théorie du vers: Rimbaud, Verlaine, Mallarmé*. Paris: Editions du Seuil.

CUTLER, A., & BUTTERFIELD, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, **31**, 218–236.

CUTLER, A., & CARTER, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, **2**, 133–142.

CUTLER, A., MEHLER, J., NORRIS, D., & SEGUI, J. (1983). A language-specific comprehension strategy. *Nature* **304**, 159–160.

CUTLER, A., MEHLER, J., NORRIS, D. G., & SEGUI, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, **25**, 385–400.

CUTLER, A., MEHLER, J., NORRIS, D., & SEGUI, J. (1989). Limits on bilingualism. *Nature*, **340**, 229–230.

CUTLER, A., MEHLER, J., NORRIS, D., & SEGUI, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology*, **24**, 381–410.

CUTLER, A., & NORRIS, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception & Performance*, **14**, 113–121.

CUTLER, A., NORRIS, D. G. & WILLIAMS, J. N. (1987). A note on the role of phonological expectations in speech segmentation. *Journal of Memory and Language*, **26**, 480–487.

DAUER, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, **11**, 51–62.

DAUER, R. M. (1987). Phonetic and phonological components of language rhythm. *Proceedings of the 11th International Congress of Phonetic Sciences, Tallinn*, **5**, 447–450.

FANT, G., KRUCKENBERG, A., & NORD, L. (1989). Rhythmical structures in text reading: A language contrasting study. *Proceedings of EUROSPEECH '89, Paris*, **1**, 498–501.

FANT, G., KRUCKENBERG, A., & NORD, L. (1991a). Durational correlates of stress in Swedish, French and English. *Journal of Phonetics*, **19**, 351–365.

FANT, G., KRUCKENBERG, A., & NORD, L. (1991b). Language specific patterns of prosodic and segmental structures in Swedish, French and English. *Proceedings of the 12th International Congress of Phonetic Sciences, Aix-en-Provence*, **4**, 118–121.

GERKEN, L. (1993). Young children's representation of prosodic structure: Evidence from English-speakers' weak syllable productions. Submitted for publication, *Journal of Memory and Language*.

HOEQUIST, C. E. (1983a). Durational correlates of linguistic rhythm categories. *Phonetica*, **40**, 19–31.

HOEQUIST, C. E. (1983b). Syllable duration in stress-, syllable- and mora-timed languages. *Phonetica*, **40**, 203–237.

JUSCZYK, P., CUTLER, A., & REDANZ, N. (1993). Infants' sensitivity to the predominant stress patterns of English words. *Child Development*, **64**.

JUSCZYK, P., & THOMPSON, E. (1978). Perception of a phonetic contrast in multisyllabic utterances by 2-month-old infants. *Perception & Psychophysics*, **23**, 105–109.

KARZON, R. G. (1985). Discrimination of polysyllabic sequences by one- to four-month-old infants. *Journal of Experimental Child Psychology*, **39**, 326–342.

KATADA, F. (1990). On the representation of moras: Evidence from a language game. *Linguistic Inquiry*, **21**, 641–646.

KUBOZONO, H. (1989). The mora and syllable structure in Japanese: Evidence from speech errors. *Language & Speech*, **32**, 249–278.

KUWABARA, H. (1982). Perception of CV-syllables isolated from Japanese connected speech. *Language & Speech*, **25**, 175–183.

MANN, V. A. (1986). Phonological awareness: The role of reading experience. *Cognition*, **24**, 65–92.

MEHLER, J., DOMMERGUES, J.-Y., FRAUENFELDER, U., & SEGUI, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, **20**, 298–305.

NORRIS, D. G. (1984). A computer-based programmable tachistoscope for nonprogrammers. *Behavior Research Methods, Instrumentation and Computers*, **16**, 25–27.

NORRIS, D. G., & CUTLER, A. (1988). The relative accessibility of phonemes and syllables. *Perception & Psychophysics*, **43**, 541–550.

OTAKE, T. (1989). Counter evidence for mora timing. *Proceedings of the Sixteenth LACUS Forum*, 313–322.

OTAKE, T. (1990). Rhythmic structure of Japanese and syllable structure. *IEICE Technical Report*, **89**, 55–61.

PETERS, A. M. (1985). Language segmentation: Operating principles for the perception and analysis of language. In D. I. Slobin (Ed.), *The Crosslinguistic Study of Language Acquisition, Vol. 2: Theoretical Issues*. Hillsdale, NJ: Erlbaum.

PORT, R. F., DALBY, J., & O'DELL, M. (1987). Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America*, **81**, 1574–1585.

SEBASTIAN-GALLES, N., DUPOUX, E., SEGUI, J., & MEHLER, J. (1992). Contrasting syllabic effects in Catalan and Spanish. *Journal of Memory and Language*, **31**, 18–32.

SEGUI, J. (1984). The syllable: A basic perceptual unit in speech processing. In H. Bouma & D. G. Bouwhuis, (Eds.), *Attention and Performance X*. Hillsdale, NJ: Erlbaum.

SEGUI, J., FRAUENFELDER, U., & MEHLER, J. (1981). Phoneme monitoring, syllable monitoring and lexical access. *British Journal of Psychology*, **72**, 471–477.

SPRING, D. R., & DALE, P. S. (1977). Discrimination of linguistic stress in early infancy. *Journal of Speech and Hearing Research*, **20**, 224–232.

WENK, B. J., & WIOLAND, F. (1982). Is French really syllable-timed? *Journal of Phonetics*, **10**, 193–216.

(Received December 29, 1991)
(Revision received June 6, 1992)