音声研究　第1巻第3号
1997(平成9)年12月
4-13頁

Journal of the Phonetic Society
of Japan. Vol.1 No.3
December 1997. pp.4–13

CONTRASTIVE STUDIES OF SPOKEN-LANGUAGE PERCEPTION

Anne CUTLER* and Takashi OTAKE**

音声言語の知覚に関する対照研究

アン カトラー*・大竹孝司**

要旨：人間が音声言語を理解する仕組みは，脳や聴覚などの観点から考えると基本的には普遍的なものと考えられなくもない。しかしながら，各言語の音韻構造が言語によって異なることから，この違いが音声言語の処理過程に反映することが十分有り得ると考えられる。この問題を理解するには，単一の言語の話者による研究もさることながら複数の言語の話者に対して同一の条件による対照研究の手法の方がより有効であることを，(1) 連続音声における分節の単位，(2) 語彙認識における超分節素の役割，(3) 音素の処理過程，(4) 心内語彙表示の音韻単位など，これまで実施した研究に基づいて論じる。

Keywords: speech perception, speech segmentation, prosody, lexical access, phonological structure

1. Introduction: Why Study Language Perception Contrastively?

The mechanisms whereby we perceive speech are the auditory system and the congnitive systems of the brain. These are much the same for all human beings. Therefore one might expect that the process of speech perception would operate in much the same manner for all people, whatever their native language. Indeed, in nearly all important ways this is so: some people have more acute hearing than others, some people are more receptive to poetry than others, some people find it easier than others to learn a second language; but these individual differences exist within every culture and do not distinguish one culture from another.

Nevertheless, there are two reasons why the process of speech perception cannot operate in a completely identical manner across all languages. One is that languages have specific features; that is, languages of the world display a wide repertoire of devices for encoding meaning in sound, and no language exhibits the full range of such features. This means that the processing of one language can draw on different characteristics of the sound signal than the processing of another, in which case the processing operations involved in listening to the two languages are inevitably not identical. We can illustrate this by considering the dimension of fundamental frequency; in some languages this dimension plays a role in encoding lexical identity, in others it does not. Thus in Cantonese the syllable /si/ means "poem" if it is spoken with a high falling tone, "time" if it spoken with a low level tone (and so on; there are six tones which could be realised on this syllable in Cantonese). In English, the syllable /si/ is also ambiguous, but the listener cannot use differences of fundamental frequency, such as between a falling or a level tone, to determine whether what has been said is *see* or *sea*. The processing operations via which lexical identity is ascertained must incorporate reference to the fundamental frequency profile of each syllable in Cantonese, but not in English.

There is however another way in which processing operations differ across languages. Recent research has shown that there exist certain strategies which facilitate linguistic processing by capitalising on the phonological structure of the language; these procedures are essentially the same across languages (i.e. universal), but because the phonological structure which they exploit differs across languages, they are realised differently (i.e. in language-specific form) across languages. One such strategy is described in

* Director, Max-Planck Institute for Psycholinguistics (マックス・プランク心理言語学研究所所長)
** Professor, Faculty of Foreign Languages, Dokkyo University (獨協大学外国語学部教授)

section 2 below; it has the effect of tailoring the perceptual system in such a way that perception of the native language is greatly assisted, but, as will be described, it actually also has an adverse side-effect, in that it exacerbates the processing difficulty offered by non-native input.

Psycholinguistics aims to study the human processing of language; but for both the above reasons this can only be achieved by undertaking contrastive studies across languages. In the case of language-specific features this is obvious; if all psycholinguistics were done in English, for instance, we would never find out anything about how human listeners can process fundamental frequency information to distinguish between words. But in the case of language-specific realisation of universal procedures, the situation is more interesting: it is only possible to find out what the universal procedure is by comparing its language-specific forms. A contrastive approach is imperative.

In this paper we summarise our joint projects in contrastive psycholinguistics from the past five years. In our work we have compared the perceptual processing of Japanese and other languages. Our studies cover a range of aspects of spoken-language processing, summarised in separate sections below: the segmentation of continuous speech, the activation of lexical forms in word recognition, phoneme identification, and the mental representation of phonetic information. In our work we have used a range of laboratory tasks, drawn both from the traditional repertoire of phonetic studies of speech perception, and from the domain of laboratory-based psycholinguistics.

2. Segmentation of continuous speech

Speech signals are continuous, yet in order to understand speech listeners must discover the constituent parts (such as words) of which each signal is composed. Studies in English and in French, conducted in the 1980s, had demonstrated a fundamental difference in the procedures which listeners applied to this segmentation problem. French listeners found it easy to segment speech input into its component syllables (Mehler, Dommergues, Frauenfelder & Segui, 1981; Segui, Frauenfelder & Mehler, 1981; Cutler, Mehler, Norris & Segui, 1986). For instance, when French listeners were asked to detect a sequence of sound such as *ba* or *bal*, they responded significantly FASTER to targets which constituted entire syllables than to non-syllabic targets (e.g. faster to *bal* than to *ba* in *balcon*, where the first syllable is *bal*, but faster to *ba* than to *bal* in *balance*, the first syllable of which is *ba*; Mehler et al., 1981). English listeners, in contrast, proved insensitive to whether or not a target constituted a syllable (responses to *ba* or *bal* were essentially the same in *balcony*, and also essentially the same in *balance*; Cutler et al., 1986). Instead, they were sensitive to the boundaries between stress units: they segmented speech at the onset of strong syllables but not at the onset of weak syllables (Cutler & Norris, 1988; Cutler & Butterfield, 1992). Segmentation errors, for instance, most often tended to consist of placing erroneous boundaries before strong syllables and overlooking boundaries before weak syllables, as when, for example, *a must to avoid* (in which only the second and last syllables are strong) is perceived as *a muscular boy* (Cutler & Butterfield, 1992).

Because in French the syllable is the basic unit of language rhythm, whereas the rhythm of English is stress-based, these studies led to the hypothesis that the segmentation of continuous speech involved a universal strategy which exploited the rhythmic structure of speech input; apparent language-specificity in processing was simply due to different implementations of the rhythmic procedure for different language rhythms. The Japanese language offered a crucial test case for the rhythmic hypothesis because it has a different kind of rhythm than the languages which had previously been tested. Our initial studies therefore undertook to test this hypothesis.

Adopting the psycholinguistic fragment detection methodology that had been used by Mehler et al. (1981) and Cutler et al. (1986), Otake, Hatano, Cutler & Mehler (1993) presented listeners with words such as *tanishi* (田螺) or *tanshi* (端子) and asked them to detect targets such as *ta* or *tan*. The results were quite different from those of the earlier studies in other languages. Firstly, there was no indication that listeners were segmenting the input into syllables. The target *ta* was detected equally rapidly and equally accurately in both types of word, although *ta* corresponds to the initial syllable of *tanishi* (田螺) but less than the initial syllable of *tanshi* (端子). In words like *tanshi* (端子), the target *ta* (which is not the initial syllable) was in fact detected significantly faster than

the target *tan* (which is the initial syllable). Thus the Japanese listeners were clearly behaving differently from the French listeners in the earlier studies. However, they were also behaving differently from English listeners; whereas English listeners detected *ba* or *bal* targets equally accurately and rapidly in words like *balance*, Japanese listeners found it very hard indeed to detect the target *tan* in *tanishi* (田螺). This is not at all surprising: *tanishi* (田螺) consists of three morae, *ta-ni-shi*, and in terms of this structure the target *tan* is simply not present in *tanishi* (田螺).[1] This result illustrated the importance of the mora in the segmentation of speech input in Japanese. Of course, the other aspects of the data tell the same story: both *tanishi* (田螺) and *tanshi* (端子) begin with the same mora *ta*, so it is again hardly surprising that responses to *ta* in both word types are equivalent.

This result provided clear support for the rhythmic hypothesis, because the mora is the basis of the ryhthmic structure in Japanese. Thus listeners appear universally to be able to exploit the rhythmic structure of speech in segmenting spoken input; but the rhythmic structure which their language provides can take different forms, so that the processing via which they implement rhythmic segmentation will be sensitive to different phonological structures. Note that the rhythmic procedure could not have been discovered on the basis of experiments in any one language alone – it was not until experiments had been conducted in several languages that the universal principle could be observed.

Subsequent experiments using the phoneme-detection task (Cutler & Otake, 1994) provided further confirmation for the importance of the mora in the segmentation of speech by Japanese listeners. In this task, listeners respond as rapidly as possible when they detect a phoneme target. Phonemes, of course, can constitute a mora by themselves: a nasal consonant in syllable-final position is moraic, and a vowel not preceded by a consonant is moraic. Indeed, Japanese listeners detect consonants and vowels significantly faster if they constitute a mora than if they do not. Thus /n/ was detected faster in *kanko* (歓呼) than in *kanoko* (鹿の子), and /o/ was detected faster in *aoki* (青木) than in *tokage* (蜥蜴). Moreover, as we will describe below, Japanese listeners also detected phonemes faster if they occurred in a moraic position in foreign-language input, which argues strongly against any suggestion that these listeners might need an orthographic representation of the input in order to perform the detection task.

The moraic nasal consonants in syllable coda position, which were used as targets here, in fact can differ in phonetic realization as a function of following context. Thus in words like *tombo* (蜻蛉) the moraic nasal is realised as the bilabial [m] before the bilabial obstruent [b], in *ringo* (林檎) it is velar before [g], in *kinri* (金利) it is alveolar before [r] and in *kondo* (今度) it is dental before [d]. Words with all four of these phonetic realizations were used as targets in a further phoneme detection experiment (Otake, Yoneyama, Cutler & van der Lugt, 1996), in which subjects were instructed to detect a sound which could be represented with the Roman character N. Responses were equally fast and accurate irrespective of effects of phonetic context (and were always faster than responses to non-moraic nasals). In contrast, Dutch listeners presented with the same materials showed clear effects of the phonetic realization; the nasal in *tombo* (蜻蛉) was responded to given an instruction to detect /m/ but not given an instruction to detect /n/, while the reverse was true of, for example, the nasal in *kondo* (今度). The results suggest that Japanese listeners' processing advantage for moraic targets involves a more abstract level of representation than the phonetic.

Finally, language-specific processing of the native language was not all that this series of studies showed. The previous English-French experiments had already shown that listening to foreign languages is rendered less efficient by the use of language-specific native procedures. Our studies provided many further demonstrations that this is so, when we compared Japanese listeners' processing of foreign languages and non-Japanese listeners' processing of Japanese. Thus French listeners who were presented with Japanese words produced a response pattern indicative of syllabic segmentation (*tan* was detected more easily than *ta* in *tanshi* (端子), but *ta* more easily than *tan* in *tanishi* (田螺), while English listeners again proved insensitive to the size of the target in both types of word (Otake et al., 1993). English listeners also showed no effects of moraic structure on phoneme detection time; /n/ was detected equally rapidly in *canopy* and *candy*, for example (Cutler & Otake, 1994). Both sets of listeners, in other words,

responded to the non-native Japanese stimuli as they did to stimuli in their native language.

As we foreshadowed above, Japanese listeners likewise showed an apparent sensitivity to moraic structure when performing phoneme detection in foreign languages. In English, for instance, they detected /n/ faster in *candy* than in *canopy* (Cutler & Otake, 1994). And they further showed response patterns consistent with mora-based segmentation when detecting word fragments such as *ba, bal,* etc. in French, in English and in Spanish (Otake, Hatano & Yoneyama, 1996). Together, this series of experiments thus suggests that speech perception processes are finely tailored to the phonological structure of the native language, and the strategies which facilitate native-language processing are difficult to inhibit when the input is in a foreign language. Where the foreign language has a phonological structure very different from that of the native language, the native strategies will of course be inappropriate. Japanese, English and French all have different rhythmic structures; this makes each language inherently difficult to listen to for anybody who has grown up as a native speaker of either of the other two.

3. Activation of lexical forms in word recognition

Languages contain tens of thousands of words, but the human articulatory and auditory systems are not suited for creating and distinguishing between that many unique completely dissimilar forms. Thus it is true of all languages that words are made up of distinguishable smaller phonetic units and, inevitably, that words resemble one another and overlap with, or are embedded within, one another. Consider the English word *key*; it is embedded in the words *keep, ski, mosquito* and many more. The Japanese word *masu* (鱒) similarly appears in *masuku* (マスク), *kamasu* (カマス), *damasukasu* (ダマスカス) and so on.

Thus a given portion of input may activate lexical forms, and word recognition is based on a process of activation of, and competition between, these forms; this can be assumed to operate in much the same manner across all languages. The input activates words which correspond to the phonetic structure perceived, and these words may compete with one another until a unique word candidate, or string of words, prevails. If the input is in fact *keep*, competition from *key* will eventually lose ground to the greater activation of *keep*. Current models of spoken-word recognition such as TRACE (McClelland & Elman, 1986) or Shortlist (Norris, 1994) are computationally explicit and implemented as computer programs which allow the simulation of results obtained in laboratory studies of spoken-word recognition. Of course, implementations of this kind require a computerized database of the entire vocabulary of a language. Such database include the publicly available CELEX detabase for English, German and Dutch (Baayen, Piepenbrock & Gulikers, 1995), or BRULEX for French (Content, Mousty & Radeau, 1990), which have allowed detailed simulation of experimental results in several languages (e.g. Norris, McQueen & Cutler, 1995; Norris, McQueen, Cutler & Butterfield, in press). To our knowledge no such publicly available complete lexical database exists for Japanese, although smaller databases are available (e.g. Sugito, 1995).

However, although word-recognition models are now much more explicit than they previously were, they are still in great need of further refinement. One important issue here is the nature of the information which constitutes the code via which stored lexical forms are activated. Most implemented models simulate only segmental information in the model's input stage; but as we pointed out in the introduction to this paper, suprasegmental information will also contribute importantly to word recognition in many languages.

In studies of word recognition in Japanese, we have addressed the issue of whether listeners draw on suprasegmental information in the early stages of word recognition. Polysyllabic words in Japanese exhibit characteristic pitch-accent patterns, and these can in fact distinguish between words which are identical in segmental structure (such as *ame* (雨) / *ame* (飴)). This may suggest that Japanese pitch accent resembles lexical tone (as, for instance, in Cantonese). To be sure, both types of structure are realised in the fundamental frequency profiles of words. But in other respects Japanese pitch accent is not at all like tone. It is essentially realised as patterns applied to polysyllabic words, and in this it more closely resembles lexical stress in English and similar languages than lexical tone in Cantonese and similar languages, since in the latter, tonal values are properties of individual syllables. This last fact means that implementing tonal information in simulating word recognition in a tone

language would be relatively straightforward: the input /si/ in Cantonese could have not one but six potential realisations, corresponding to the six possible tones. Stress and pitch accent, on the other hand, lend themselves much less well to this quasi-segmental method of implementation. There is evidence from experiments in English that the suprasegmental correlates of stress do not participate in the initial activation of lexical candidates (Cutler, 1986), and it is widely assumed that Japanese pitch accent serves less to perform a lexical function (i.e. distinguish between words) than a syntactic one (i.e. mark a word boundary) or even a social one (i.e. convey a dialectal affiliation; Kindaichi, 1967; Vance, 1987). Thus pitch accent constitutes an interesting case for examining the input level to lexical activation.

In our experiments addressing this issue we made use both of phonetic techniques (categorisation of syllables) and psycholinguistic experimental methods (gating, i.e. the assessment of the contribution of word fragments to word recognition, and repetition priming in lexical decision, a reaction time paradigm). Our first experiment (Cutler & Otake, 1996) simply asked whether listeners can extract from naturally spoken Japanese words reliable pitch accent information which could narrow the set of possible word candidates. We recorded a set of words with the segmental structure CVCV, all containing the syllable *ka*. Half of the words had HL accent pattern, half LH: *kage* (影) / *kagi* (鍵), *baka* (馬鹿) / *gaka* (画家). For each pattern, in half of the words the syllable *ka* was word-initial, in half word-final. The *ka* syllables were extracted from each production and presented to listeners who chose for each token between two words from which it might have come. The results showed a very high proportion of correct responses. Identification was more accurate for H than for L syllables and for initial than for final syllables. This suggests that pitch accent information is realized most clearly in just the position where it would be of most use for listeners in on-line spoken-word recognition. Acoustic analyses of the stimuli showed that H and L syllables differed principally in fundamental frequency (higher in H syllables; more varying in L syllables), and correlations confirmed that listeners' judgements were based on these acoustic effects.

In our next experiment (Otake & Cutler, 1997) we directly followed up the implication that pitch accent information can be used early in the recognition of a word. In a gating study we examined the recognition of pairs of Japanese words such as *nimotsu* (荷物) / *nimono* (煮物), beginning with the same bimoraic CVCV sequence but with the accent pattern of this initial CVCV being HL in one word and LH in the other. These words were presented, in increasingly large fragments, to Japanese listeners, who recorded a guess after each fragment as to the word's identity, and a confidence rating for that guess. The results showed that the accent patterns of the word guesses corresponded to the accent patterns of the actually spoken words with a probability significantly above chance from the second fragment onwards – i.e., from the middle of the vowel in the first mora of the word. Accent correspondence averaged 79.6% at this point, rising to 89% by the fourth fragment (vowel of second mora). Moreover, the confidence which listeners had in their (erroneous) guesses was also significantly higher when the guessed word began with the correct accent profile (HL- or LH-) of the actually spoken word, than when the guessed word had differing accent pattern. This demonstrates that Japanese listeners can and do exploit pitch-accent information effectively at an early stage in the presentation of a word, and use it, where possible, to constrain selection of lexical candidates (Of course, there are many cases of ambiguity – such as *kumo* (雲) / *kumo* (蜘蛛) – where pitch-accent information will not resolve the choice between meanings. In English, similarly, stress can distinguish *trusty* from *trustee*, but it cannot distinguish between *bridal* and *bridle*, which are pronounced identically.).

Fianlly, a reaction-time experiment (Cutler & Otake, forthcoming) confirmed that pitch-accent information constrains lexical activation in a rapid and effective manner. This experiment exploited the repetition priming effect in lexical decision: response time to decide that a given input does or does not correspond to an existing word is significantly shorter if the word has already been presented once in the same experiment. It further drew on the existence in Japanese of minimal pitch-accent pairs such as *ame* (雨) / *ame* (飴). If listeners hear *ame* (雨) HL, and then later hear *ame* (雨) HL again, it is to be expected that their response on the second occurrence will be faster than their response on the first occurrence. But what if they hear *ame* (雨) HL after previously hearing *ame* (飴)

LH? Will *ame* (飴) LH produce any partial activation of *ame* (雨) HL? Our experiments showed that there was no facilitation at all of responses to a word which occurred after an earlier presentation of its minimal pitch-accent pair. Only a second presentation of the same word produced facilitation. Again, this is exactly the result that one would expect if listeners are effectively exploiting pitch-accent information in word activation.

This series of experiments suggests that the input to the activation stage of word recognition incorporates suprasegmental information, not just in languages where this information is encoded in a segmental-like manner, but also where a higher level of prosodic organisation is involved. Recent evidence suggests that stress information can also contribute to lexical activation in Dutch (Koster & Cutler, 1997; van Donselaar & Cutler, 1997). If human listeners exploit such information at an early stage of word processing, it is incumbent upon the developers of word-recognition models to adapt their computational models so that in this respect also they can simulate the human performance. Again, cross-linguistic studies have been necessary to demonstrate the importance of this aspect of processing in word activation.

4. Phonemic processing

The phoneme is the smallest structural unit in terms of which speech can be sequentially described. Many models of recognition, as was pointed out in section 3 above, assume a segmental interpretation of the input. Often this is purely for computational convenience; but there are also a number of theoretical approaches which maintain that the perception of speech necessarily involves identification of phonemes in sequences. Irrespective of whether explicit representation of phonemes is required for speech recognition, however, many interesting issues arise in the consideration of phonemic processing.

Many of these issues can, again, only be properly addressed via a contrastive approach. For example, the phonemic level interacts with other levels of description. Thus in Japanese, as we pointed out in section 2 above, some phonemes are morae and some not, and this has proved crucial for the interpretation of the results of the phoneme detection studies described in that section. Moreover, as was described in that section, this interaction of the phonemic with the moraic level of structure in Japanese affects Japanese listeners' processing not only of their native language but also of foreign-language input; in contrast, non-native listeners do not process Japanese input the way native listeners do.

Similarly, phonemes come in different kinds – vowels and consonants. This has important processing consequences in some languages; English listeners, for example, process vowels in certain laboratory tasks more cautiously than consonants (van Ooijen, 1994, 1996). Thus phoneme detection times are in general longer for English vowels than for consonants (van Ooijen, 1994). Cutler and Otake (1994) replicated this finding in their phoneme detection study, and demonstrated moreover that it was yet another case in which native procedures are applied to non-native input; the English listeners also showed slower responses to Japanese vowels than to Japanese consonants, even though native Japanese listeners showed no such effects with the same input. This latter finding, of course, further demonstrated that the vowel-consonant detection difference is not a language-universal phonomenon.

Furthermore, some phonemes adopt a different form depending on the surrounding phonetic context, and how such phonetic transformations are processed raises important theoretical issues. As described in section 2, Otake et al. (1996) found that Japanese listeners responded equally accurately and rapidly to different phonetic realisations of an abstractly specified nasal target. This study however also explored the effects of mismatch between the place of articulation of a moraic nasal and of a following stop consonant. The nasal consonants described above (e.g. the nasals in *tombo* (蜻蛉), *ringo* (林檎), *kondo* (今度), *kinri* (金利) were cross-spliced across place of articulation contexts. The Japanese listeners still showed no significant place of articulation effects, suggesting that these listeners are capable of very rapid abstraction from phonetic realisation to a unitary representation of moraic nasals. Responses were, however, faster and more accurate to unspliced than to cross-spliced nasals. Moreover, when Japanese listeners were asked to detect the phoneme following the (cross-spliced) moraic nasal, their responses were adversely affected by mismatch between nasal and context. Dutch control subjects, presented with the same materials, did not show such adverse effects.

This result could be interpreted as suggesting that the Japanese listeners can use the phonetic realisation of a moraic nasal effectively to obtain anticipatory information about following phonemes. The Dutch listeners, on the other hand, could be held to be unable to use the potential anticipatory information, at least with foreign input. However, some caution is required here. Phonological rules such as assimilation of place of articulation are not implemented identically in all languages. In particular, the way that they are implemented differs in Japanese and in Dutch. In Japanese, the rule is obligatory, i.e. it admits of no exceptions. In other languages the same rule exists but in an optional form. English is one such langauge; in English, it is possible to realise *sunbathing* as *sumbathing*, but equally possible to produce the unassimilated form. Dutch, in this instance, resembles English.

To explore the differences in the use of the assimilation rule in Dutch and Japanese, we compared the perception of assimilated forms by speakers of these two languages using a word-blending task (Otake and Cutler (forthcoming)). Subjects were asked to construct a blend of two pseudo-names, using the first part of the first name and the last part of the last name. For instance, Japanese subjects were asked to blend *ranga* with *serupa*, or *kumba* with *soroki*. Japanese were predicted to perceive the nasals as unmarked for place of articulation and to produce assimilated nasals in the blends (e.g. *rampa, kunki*); if the nasals were represented in its surface form, on the other hand, this would be preserved in the blends (*ranpa, kumki*). Japanese subjects indeed produced forms consistent with unmarked underlying-form representation. When analogous Dutch materials (*mengkerk / trabeek*; *stambest / sliekoop*) were presented to Dutch subjects, however, they produced forms consistent with surface representations (*mengbeek*; *stamkoop*) rather than with underlying representations (*membeek*; *stangkoop*). In subsequent experiments the Japanese subjects were presented with the Dutch materials, and vice versa. The crosslinguistic task proved inordinately difficult for both subject populations, but those responses which could be analysed showed that the Japanese listeners still used an underlying representation and the Dutch listeners a surface representation. This experiment confirmed that the place-of-articulation assimilation rule, despite its similarity of application, has fundamentally different status in Japanese and in Dutch.

This is important for interpretation of the phoneme detection results, for the following reason. In a language like Japanese, in which the assimilation is obligatory, it is not in fact possible to test the question of whether the presence of assimilation facilitates processing of the second of the two phonemes. A clean test of this question requires that the same word be presented both with and without assimilation, in an otherwise identical experimental situation. But if the assimilation rule is obligatory, the unassimilated case is not an available option – or rather, the experimental situation is no longer identical, because one form of the word is acceptable and the other unacceptable. Thus a difference in processing could reflect inhibition due to the rule violation just as easily as it could reflect facilitation due to the rule's application; the two explanations can never be teased apart.

The contribution of assimilation to phoneme processing can in fact only be understood by contrastive studies, in which cases of obligatory rules (such as place assimilation in Japanese) are compared with optional rules (such as the same place assimilation in Dutch or English). Where the rule is optional, equally acceptable input can be presented with and without assimilation, and it can be determined whether the assimilation leads to facilitated processing. Note, however, that an optional rule does not allow a test of whether rule violation causes inhibition of processing, because there is no such thing as a violation if the rule is optional. Thus the complete picture can never be obtained from experiments in any one language alone, but only from contrastive experiments across languages.

In fact, studies in English (Koster, 1987; Gaskell & Marslen-Wilson, 1996, in press) and in Dutch (Koster, 1987; Kuijpers & van Donselaar, submitted) combine to suggest that there is no facilitatory effect of assimilation; phoneme detection for stop consonants is unaffected by whether or not assimilation has applied to an immediately preceding consonant. But a violation always adversely affects responses. This was demonstrated in the Dutch experiments of Kuijpers and van Donselaar, which concerned not place assimilation of voicing. Thus the Dutch word *kaas* "cheese", ending in /s/, may be pronounced *kaaz* in the word *kaasboer* "cheesemonger" because the following syl-

lable begins with the voiced sound /b/ and the /s/ can assimilate in voicing to the /b/. This is a particularly interesting case because it in fact triggers violation of a separate rule, in that it causes a syllable to end with a voiced obstruent, in contravention of the Dutch syllable-final devoicing rule, which is otherwise obligatory. In keeping with the general failure to find facilitation, Kuijpers and van Donselaar found that the /b/ in *kaasboer* was detected just as rapidly whether the word was pronounced *kaasboer* or *kaazboer*. But the /p/ in *kaasplank* "cheese board" was detected significantly more slowly when the word was incorrectly pronounced *kaazplank* than when it was correctly pronounced *kaasplank*; violation of an obligatory rule adversely affected phonemic processing.

This contrastive set of data across languages allows us to construct a coherent picture of the role of assimilation in phonemic processing. In the Japanese experiment of Otake et al. (1996), we can now see that the response time difference reflected not facilitation but inhibition. Violation of the rule caused slower responses from those listeners who commanded the Japanese assimilation rule, i.e. the Japanese listeners. The Dutch listeners, on the other hand, for whom the rule is optional, showed no adverse affect of the violation, and also no facilitatory effect of the presence of assimilation – fully consistent with their response pattern in their own language. In other words, the processing consequences of assimilation phenomena can only be understood as long as data are compared across languages with obligatory and optional assimilation.

5. Mental representation of phonetic information

In another set of studies we considered issues of mental representation – the stored forms of linguistic knowledge which listeners draw on in processing, and the metalinguistic knowledge which they have about the sound patterns of their native language. Issues of representation, and in particular issues of metalinguistic knowledge, are in principle separate from issues of processing, though in fact there are many points of contact between the two, and obviously both are part of the understanding of language use which psycholinguists seek to achieve.

Linked with the above questions about phonemes, for instance, is the issue of how aware listeners are of the sequential makeup of speech input. Awareness of phonemes is facilitated by learning to read in an alphabetic orthography. But if – as in Japanese – one's native language orthography is not alphabetic, then the phoneme may not be an easy unit to manipulate. In fact some parts of our research which have already been described also shed light on such issues. Thus the experiments of Cutler and Otake (1994) and Otake et al. (1996) clearly show that phoneme detection is possible for Japanese listeners even when the phoneme is non-moraic. Detection of moraic phonemes is faster, but non-moraic phonemes can still be detected; listeners do respond to /n/ in *kanoko* (鹿の子) and to /o/ in *tokage* (蜥蜴). Nevertheless other experiments show that conscious awareness of the actual sequence of phonemes in Japanese can be low. Thus Japanese listeners have difficulty in identifying reliably an intervocalic sequence of two consonants, such as /apto/ or /atko/ (whereas Dutch listeners experience no such problem with the same identification task; Kekehi, Kato & Kashino, 1996). Japanese listeners likewise have difficulty distinguishing such a VCCV sequence from the same sequence with a vowel inserted between the two consonants, e.g. /agmi/ from /agumi/ (Dupoux, Kakehi, Hirose, Pallier & Mehler, submitted). This suggests that awareness as reflected in explicit judgements, and ability to process in a speeded laboratory task requiring simply a detection response, are not necessarily isomorphous.

In the area of segmentation we have other, separate evidence for this separation. Otake, Davis and Cutler (1995) presented Japanese, British English and American English listeners with spoken polysyllabic words, of a variety of different structures, in their native language; the listeners were asked to mark on a written transcript of each word the first natural division point in the word. The results showed clear and strong patterns of consensus, indicating that listeners have available to them conscious representations of within-word structure. In fact, the most common response across all word structures for both groups of English listeners was the same: a consonant-vowel-consonant (CVC) sequence, such as *can* in *canopy, cancel* or *canteen*. Exactly this CVC response was also the most common choice of Japanese listeners for words with a nasal or geminate coda to the initial syllable, such as *kenri* (権利) or *katto* (カット).

Only in one case did one listener group fail to prefer a CVC as the first natural within-word unit: Japanese listeners preferred to divide words which began with CVCV sequence either after the first CV of after the CVCV string; thus *kamera* (カメラ) could be divided ka/mera or kame/ra.

Orthography did not play a strongly deciding role in the results. There was, for instance, no significantly greater preference to place a segmentation boundary after the initial CV of words which were presented in kana orthography than of words which were presented in kanji characters. Instead, we proposed that the results reflected listeners' appreciation of within-word structure.

A comparison with the research described in section 2 above shows clearly that the patterns of response in this explicit segmentation task were at variance with results from on-line studies of speech segmentation. In the earlier studies, a CVC sequence formed a less easily detectable target for Japanese listeners than a CV sequence, in Japanese words like *tanshi* (端子) or *kinri* (金利). Likewise, English listeners did not show a detection advantage for CVC over CV targets in either of the two word types tested in the earlier experiments. Here, however, when subjects are required to make explicit segmentation judgements, CVC is the preferred segmentation in both these cases. This suggests that the explicit segmentation task may tap not those representations used in on-line listening, but levels of representation which involve much richer knowledge of word-internal structure.

6. Conclusion

Together these studies have illuminated several facets of the complex process of speech perception. Our work shows that language-specificity in processing exists at many levels. We have concentrated on studies in which Japanese is compared with other languages; but of course a myriad other comparisons would be equally possible and instructive. Just these comparisons, however, have provided crucial support for the rtythmic hypothesis in speech segmentation; have allowed us to see that basing the input to the activation stage of word recognition models on segmental information alone is an over-simplification, since human listeners use suprasegmental information also; have illuminated the role of phonetic assimilation processes in spoken-word processing; and have led to further understanding of the levels of processing involved in differing types of psycholinguistic task.

Not only are there many futher language contrasts which could provide fruitful insight into the course of human spoken-language perception, there are also many other aspects of speech perception which have not yet been considered from a cross-linguistic viewpoint. Many aspects of our current speech perception theories, for instance, may rest on too narrow an evidential base, so that they are unintensionally biased towards a particular language or familiy of languages; the availability of cross-linguistic evidence could in such a case lead to greater theoretical precision. Contrastive studies of spoken-language perception have much to offer!

NOTE

1) The importance of the mora in segmenting Japanese is, of course, reinforced by the kana orthographies. Note, however, that in these experiments, the subjects were told what to listen for either via auditory presentation or via a visual presentation in Roman letters; never were they instructed via the use of kana. For further discussion of the role of orthography in this speeded detection task, see Otake et al. (1993).

REFERENCES

Baayen, R. H., Piepenbrock, R. & Gulikers, L. (1995) *The CELEC Lexical Database* (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Content, A., Mousty, P. & Radeau, M. (1990) "BRULEX": Une base de données lexicales informatisée pour le français écrit et parlé," *Année Psychologique*, 90, 551–556.

Cutler, A. (1986) "*Forbear* is a homophone: lexical prosody does not constrain lexical access." *Language and Speech*, 29, 201–220.

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

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

Cutler, A. & Otake, T. (1994) "Mora or phoneme? Further evidence for language-specific listening," *Journal of Memory and Language*., 33, 824–844.

Cutler, A. & Otake T. (1996) "The processing of word prosody in Japanese," *Proceedings of the 6th Australian*

International Conference on Speech Science and Technology, Adelaide; 599–604.

Cutler, A. & Otake, T. (forthcoming) "Pitch accent in spoken-word recognition in Japanese."

van Donselaar, W. & Cutler, A. (1997) "Exploitation of sress information in spoken-word recognition in Dutch," Paper presented to the 134th Meeting, Acoustical Society of America, San Diego, December.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C. & Mehler, J. (submitted) "Epenthetic vowels in Japanese: A perceptual illusion?"

Gaskell, M. G. & Marslen-Wilson, W. D. (1996) "Phonological variation and inference in lexical access," *Journal of Experimental Psychology: Human Perception and Performance*, 22, 144–158.

Gaskell, M. G. & Marslen-Wilson, W. D. (in press) "Mechanisms of phonological inference," *Journal of Experimental Psychology: Human Perception and Performance*.

Kakehi, K., Kato, K. & Kashino, M. (1996) "Phoneme/syllable perception and the temporal structure of speech," In T. Otake & A. Cutler (Eds.) *Phonological Structure and Language Processing: Cross-Linguistic Studies*, 125–143. Berlin: Mouton de Gruyter.

Kindaichi, H. (1967) "Nihongo no akusento," *Nihongo Onin no Kenkyu*, 198–230. Tokyo: Tokyodoo Shuppan.

Koster, C. J. (1987) *Word Recognition in Foreign and Native Language*. Dordrecht: Foris.

Koster, M. & Cutler, A. (1997) "Segmental and suprasegmental contribution to spoken-word recognition in Dutch," *Proceedings of EUROSPEECH 97*, Rhodes; 2167–2170.

Kuijpers, C. T. L. & van Donselaar, W. (submitted) "Phonological variation and phoneme identification in Dutch."

McClelland, J. L. & Elman, J. L. (1986) "The TRACE model of speech perception," *Cognitive Psychology*, 18, 1–86.

Mehler, J., Dommergues, J.-Y., Frauenfelder, U. & Segui, J. (1981) "The syllable's role in speech segmentation," *Jounal of Verval Learning & Verbal Behavior*, 20, 298–305.

Norris, D. G. (1994) "Shortlist: A connectionist model of continuous speech recognition," *Cognition*, 52, 189–234.

Norris, D. G., McQueen, J. M. & Cutler, A. (1995) "Competition and segmentation in spoken word recognition," *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21, 1209–1228.

Norris, D. G., McQueen, J. M., Cutler, A. & Butterfield, S. (in press) "The possible-word constraint in the segmentation of continuous speech," *Cognitive Psychology*.

van Ooijen, B. (1994) *Processing of Vowels and Consonants*. PhD Dissertation, University of Leiden.

van Ooijen, B. (1996) "Vowel mutability and lexical selection in English: Evidence from a word reconstruction task," *Memory & Cognition*, 24, 573–583.

Otake, T. & Cutler, A. (1997) "Early use of pitch accent in Japanese spoken-word recognition," Paper presented to the 134th Meeting, Acoustical Society of America, San Diego, December.

Otake & Cutler (forthcoming) "Obligatory and optional place assimilation: A cross-linguistic production study."

Otake, T., Davis, S. & Cutler, A. (1995) "Listeners' representations of within-word structure: A cross-linguistic and cross-dialectal investigation," *Proceedings of EUROSPEECH 95*, Madrid; Vol. 3, 1703–1706.

Otake, T., Hatano, G., Cutler, A. & Mehler, J. (1993) "Mora or syllable? Speech segmentation in Japanese," *Journal of Memory and Language*, 32, 258–278.

Otake, T., Hatano, G. & Yoneyama, K. (1996) "Speech segmentation by Japanese listeners," In T. Otake & A. Cutler (Eds.) *Phonological Structure and Language Processing: Cross-Linguistic Studies*, 183–201. Berlin: Mouton de Gruyter.

Otake, T., Yoneyama, K., Cutler, A. & van der Lugt, A. (1996) "The representation of Japanese moraic nasals," *Journal of the Acoustical Society of America*, 100, 3831–3842.

Segui, J., Frauenfelder, U. H. & Mehler, J. (1981) "Phoneme monitoring, syllable monitoring and lexical access," *British Journal of Psychology*, 72, 471–477.

Sugito, M. (1995) *Osaka/Tokyo Akusento Onsei Jiten* CD-ROM. Tokyo: Maruzen.

Vance, T. J. (1987) *An Introduction to Japanese Phonology*. Albany: State University of New York Press.

投稿規定の改訂について

本年第2回編集委員会において，投稿規定の改正について審議をし，原稿の長さを次のように改訂することにしました。

原稿の長さは，組み上がり10頁以内を原則とし，それを超える分については，著者の実費負担とする。

この規定は，今回の12月号から適用いたします。なお，超過分については，(図表等により多少の違いが生じますが) 1頁およそ1万円くらいになる予定です。

(編集委員長　原口庄輔)