ELSEVIER

Speech Communication 21 (1997) 3–15

SPEECH COMMUNICATION

The comparative perspective on spoken-language processing

Anne Cutler *

Max-Planck-Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands

Received 12 November 1996; accepted 18 November 1996

Abstract

Psycholinguists strive to construct a model of human language processing in general. But this does not imply that they should confine their research to universal aspects of linguistic structure, and avoid research on language-specific phenomena. First, even universal characteristics of language structure can only be accurately observed cross-linguistically. This point is illustrated here by research on the role of the syllable in spoken-word recognition, on the perceptual processing of vowels versus consonants, and on the contribution of phonetic assimilation phonemena to phoneme identification. In each case, it is only by looking at the pattern of effects across languages that it is possible to understand the general principle. Second, language-specific processing can certainly shed light on the universal model of language comprehension. This second point is illustrated by studies of the exploitation of vowel harmony in the lexical segmentation of Finnish, of the recognition of Dutch words with and without vowel epenthesis, and of the contribution of different kinds of lexical prosodic structure (tone, pitch accent, stress) to the initial activation of candidate words in lexical access. In each case, aspects of the universal processing model are revealed by analysis of these language-specific effects. In short, the study of spoken-language processing by human listeners requires cross-linguistic comparison.

Résumé

Les psycholinguistes s'efforcent de construire un modèle du traitement de langage humain en général. Mais ceci n'implique pas qu'ils doivent limiter leurs recherches aux aspects universels de la structure linguistique et éviter les recherches sur les phénomènes spécifiques à telle ou telle langue. Tout d'abord, mêmes les caractéristiques universelles de la structure du langage ne peuvent être observées de façon précise qu'en étudiant plusieurs langues. Ce point est illustré dans cet article par des travaux sur le rôle de la syllabe dans la reconnaissance de mots parlés, sur le traitement perceptif des voyelles par rapport aux consonnes et sur la contribution des phénomènes d'assimilation phonétique à l'identification des phonèmes. Dans chaque cas, ce n'est qu'en étudiant les phénomènes de façon comparative entre les langues que l'on peut en comprendre le principe général. D'autre part, les traitements spécifiques à chaque langue peuvent certainement aider à identifier un modèle universel de la compréhension du langage. On illustre ici ce second point par des études sur l'exploitation de l'harmonie vocalique dans le segmentation lexicale en Finois, sur la reconnaissance de mots en Néerlandais avec et sans épenthèse et sur la contribution des différents types de structure prosodique lexicale (ton, accent intonatif, accent lexical) à l'activation initiale des mots candidats dans l'accès au lexique. Dans chaque cas, les aspects du modèle

* E-mail: anne.cutler@mpi.nl.

PII S0167-6393(96)00075-1

universel de traitement sont révélés par l'analyse d'effets spécifiques à chaque langue. En résumé, l'étude du traitement du langage parlé par l'être humain requiert des comparaisons inter-langues.

1. Introduction

If everyone in the world spoke the same language, we would have been unable to achieve much of the knowledge we have of the processes underlying spoken-language processing by human listeners. Psycholinguists working in this area attempt to model the listening process, and their aim is a model which holds for all human listeners and is therefore valid irrespective of the language which is being listened to. Separate models accounting for the processing of English, of Japanese, of Sesotho and so on might perhaps be warranted if we were to believe that each different language represented a fundamentally different achievement of the human mind. However, we are quite certain that this is not the case. Humans are born equipped to acquire language per se, not a particular language; whatever language is spoken in the environment will be acquired equally well. Thus the model of spoken-language processing which we aim to construct must be a universal one; and the purpose of this contribution is to argue that such a universal model can only be arrived at via comparative studies across languages.

This belief has not always been held in psycholinguistics. In an earlier period of psycholinguistic research (see (Cutler, 1985) for references) it appeared to be frequently assumed that the quest for a universal model could best be served by studying only the processing of universal aspects of linguistic structure, which, because they were evident in all languages, could be studied in any language. Language-specific effects, evident in only few languages, could best be left until the nature of the universal model had been satisfactorily established.

There is probably no psycholinguist who would subscribe to this assumption today, which is fortunate, because it is wrong on both counts. As the following sections will show (a) the processing of universal aspects of linguistic structure cannot be understood without comparative research, and (b) language-specific effects can prove highly informative in our search for the universal model. Each of these arguments will be illustrated with three pieces of evidence, all of which come from my own research area, namely the recognition, by human listeners, of sounds and words in spoken language.

2. Universal

2.1. Syllables

In the days in which psycholinguistics was effectively monolingual, the psycholinguist's single language was virtually always English. Thus it is perhaps instructive to focus in part on the question of whether results of psycholinguistic experiments in English actually generalise to other languages. The first case study concerns a question which has in fact been asked rather frequently in the last few decades of spoken-word recognition research, namely: do syllables play a role in the recognition of spoken words?

In order to answer this apparently simple question in a psycholinguistic laboratory we simplify it still further – the art of psycholinguistics, as of any experimental science, being to generate straightforward questions that may be addressed by straightforward experiments. In spoken word recognition research this often results in an experimental situation in which subjects listen to words or sentences and press a button as soon as they detect an occurrence of a specified target, whereupon their response time (RT) from actual occurrence of the target to button-press is measured. The question about the syllable may thus be addressed via such a detection task; do subjects respond faster to targets which exactly correspond to syllables than to targets which are not exact syllables?

The question of exactly what corresponds to a syllable in English is not always uncontroversial, but some clear cases do exist. In the word *balcony* for example the first vowel /ae/ is followed by two consonants /l/, /k/ which cannot together form a syllabic onset – no English syllable begins /lk/. Thus there cannot possibly be a syllable boundary immediately before these two consonants, and the first vowel in the word *balcony* cannot possibly be

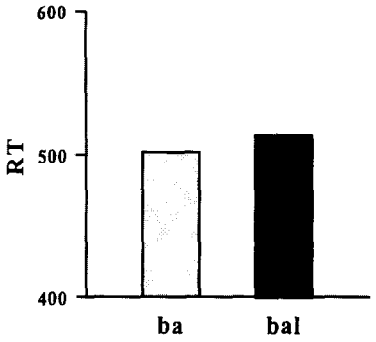

Fig. 1. Fragment detection response times (msec) of English listeners to CV (e.g. *ba*) and CVC (e.g. *bal*) targets in English words with closed initial syllables (e.g. *balcony*).

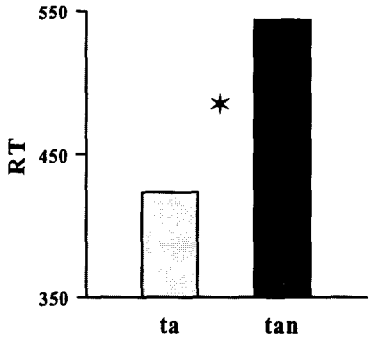

Fig. 2. Fragment detection response times (msec) of Japanese listeners to CV (e.g. *ta*) and CVC (e.g. *tan*) targets in Japanese words with closed initial syllables (e.g. *tanshi*).

the end of a syllable. A target such as *ba* could in consequence never be a syllable in *balcony*; the target *bal*, on the other hand, can be a syllable of *balcony*. The subjects' responses (depicted in Fig. 1) showed that the two targets *ba* and *bal* were equally easy, or equally difficult (Cutler et al., 1986), motivating the conclusion that the syllable as such plays no role in English listeners' processing of words like *balcony*.

Should we expect to find the same result had this experiment been conducted not in English but in some randomly chosen other language? Fortunately the experiment has been done in other languages (which in fact were not randomly chosen ones). Two of these are Japanese and French; both of these languages also have consonant combinations which cannot possibly constitute a syllabic onset but nevertheless do occur in sequence within words. In Japanese, for example, the consonants [n] [ʃ] cannot constitute an onset, so that the word *tanshi* once again cannot possibly have a syllable boundary after the first vowel, though it may have one between the [n] and the [ʃ]. In French, as in English, no syllable could begin with [lk], so that the word *balcon* could not contain a boundary before the [l], but only after it. Nevertheless, when the same experiment was conducted in these two languages, the results proved completely different from the English result. (The star in Figs. 2 and 3, and in later figures, denotes a difference that is statistically significant.) Both in Japanese and in French a significant difference was observed between targets which exactly corresponded with a syllable and targets which did not. Unfortunately it was not the same difference: in Japanese (see Fig. 2), the subjects responded to the syllabic targets significantly more slowly than to the non-syllabic targets (e.g. *tan* versus *ta* in *tanshi*; Otake et al., 1993). The French subjects (see Fig. 3) responded significantly FASTER to the syllabic than to the non-syllabic targets (e.g. *bal* compared with *ba* in *balcon*; Mehler et al., 1981).

In other words, a question which has universal application produces from experiments in three different languages three different answers: the shorter target is easier, the longer target is easier, or the two are equivalent. This is hardly supportive of the proposal that a universal construct such as the syllable should elicit the same pattern of effects in any language.

This is a highly simplified excerpt from a long series of studies; more details may be found, for

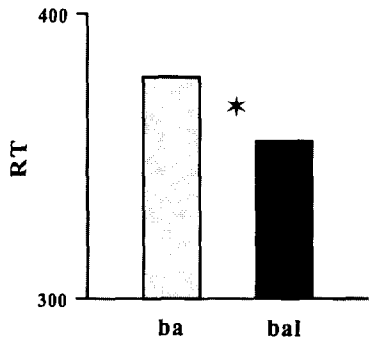

Fig. 3. Fragment detection response times (msec) of French listeners to CV (e.g. *ba*) and CVC (e.g. *bal*) targets in French words with closed initial syllables (e.g. *balcon*).

instance, in (Mehler et al., 1996; Otake et al., 1996a). The remainder of this contribution is devoted to much more recent work, but before leaving the topic of the syllable here is a brief summary of the conclusions to which this research led. The processing of phonological constructs, even universally applicable constructs such as the syllable, depends on their role in the phonological structure of the language in question. The syllable's role in the phonological structure of French includes being the basic unit of rhythm in that language (for example in French poetry). In Japanese the basic unit of rhythm is something else, namely the mora. The mora is also a universally applicable phonological construct; it is a subsyllabic unit, and for present purposes the most relevant information is that a short open syllable such as *ta* is a single mora, whereas a closed syllable such as *tan* must contain at least two morae. The experiments described in this section therefore in effect contrasted moraic with nonmoraic targets, as well as syllabic with nonsyllabic. The importance of the mora in the processing of Japanese, and of the syllable in the processing of French, have been repeatedly established in many experiments with many different methodologies (Mehler et al., 1996; Otake et al., 1996a). The importance for processing presumably reflects the role that these constructs play in the two languages' rhythm. The fact that neither syllable nor mora seems at an advantage in the processing of English similarly reflects the fact that neither syllable nor mora plays a role in the rhythm of English, since English rhythm is based on stress units.

The whole series of experiments on syllables, morae and stress enabled my colleagues and myself to propose a universal explanation of the role of language rhythm in the recognition of spoken language, namely that the segmentation of continuous speech into individual words exploits rhythmic structure. But even though our explanation involved concepts which are truly universal and hence can be applied to any language, such as syllable and mora, it could not have been arrived at on the basis of experiments in any one language alone – it was not until we had done experiments in several languages that we could see that there was a univeral principle operating.

The following two further case studies, both from more recent work in my lab, similarly show how even universal constructs must really be investigated in more than one language.

2.2. Vowels and consonants

Detection tasks, in which subjects listen to speech and respond as soon as a specified target appears, enabled us to discover the processing roles of syllables and morae. The simplest task of this kind is phoneme detection, in which the target is a single speech sound, a phoneme. Phoneme detection with English listeners has recently produced the somewhat surprising finding that not all phoneme targets are equivalently easy: vowels are much harder detection targets than consonants (Van Ooijen, 1994; Cutler et al., 1996). For instance, RT is longer to vowels than to nasal consonants, whether English listeners are listening to speech in their own language (Fig. 4)

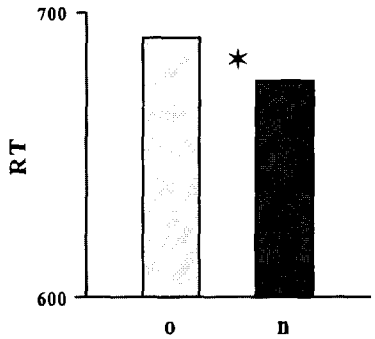

Fig. 4. Phoneme detection response times (msec) of English listeners for vowel and consonant targets in English words.

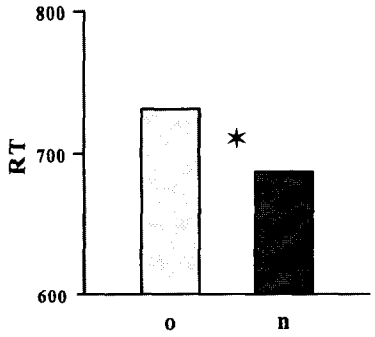

Fig. 5. Phoneme detection response times (msec) of English listeners for vowel and consonant targets in Japanese words.

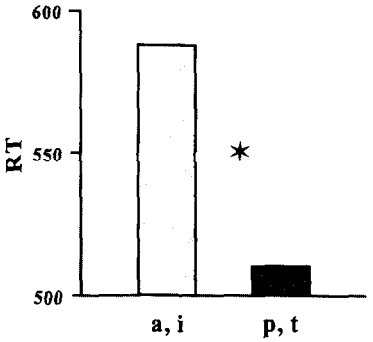

Fig. 6. Phoneme detection response times (msec) of English listeners for distinct vowel and confusible stop consonant targets in English words.

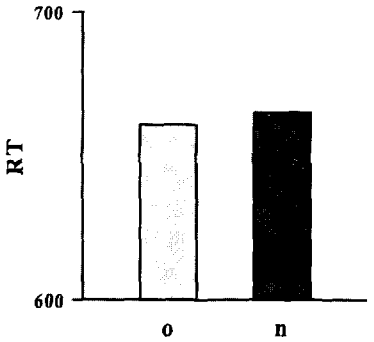

Fig. 8. Phoneme detection response times (msec) of Japanese listeners listeners for vowel and consonant targets in Japanese words.

or to Japanese words (Fig. 5; Cutler and Otake, 1994). Further, English listeners' RTs are significantly longer to vowels (even vowels which are themselves very easily distinguishable e.g. /a/, /i/) than to consonants which in principle are themselves fairly difficult such as the confusible stop consonants /p and /t/ (Fig. 6), or fricatives such as /s/ and /v/ (Fig. 7; Van Ooijen et al., Forthcoming).

No language anywhere consists solely of vowels or solely of consonants; so should we expect this result to be equally valid for every language in the world? Needless to say, this proves not to be the case. If Japanese subjects are presented with those same Japanese words we do NOT find the RT difference that English listeners showed when they heard those items (Fig. 8; Cutler and Otake, 1994). And when Spanish listeners perform the same phoneme detection task with Spanish words, they fail to show a significant difference in RT to vowels versus stop consonants (Fig. 9), and they show a difference in the OPPOSITE direction with vowels versus fricatives (Fig. 10) – for Spanish listeners, fricatives seem to be the hardest type of target to detect (Van Ooijen et al., Forthcoming).

It would therefore have been inappropriate to have concluded from the English finding, robust though it certainly seems, that vowels are in general harder to detect than consonants. In at least two other languages the situation is different. There are several reasons for this. First, there is the simple fact of vowel repertoire size – both Japanese and Spanish are five-vowel systems, while English has an extremely crowded vowel space, so that any one English vowel has more neighbours with whom it can be confused, and is thus in principle less distinctive. This cannot however be the whole explanation, since

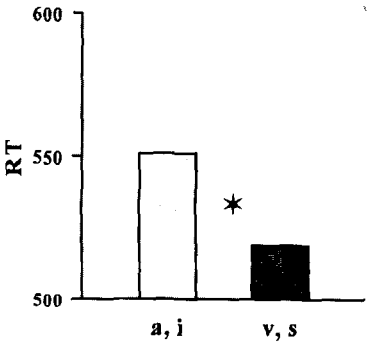

Fig. 7. Phoneme detection response times (msec) of English listeners for vowel and fricative targets in English words.

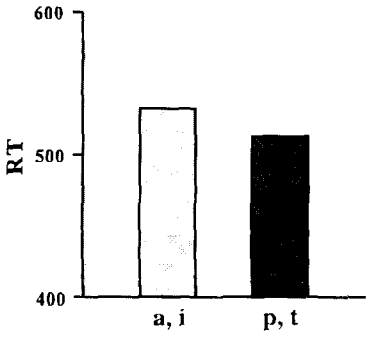

Fig. 9. Phoneme detection response times (msec) of Spanish listeners for distinct vowel and confusible stop consonant targets in Spanish words.

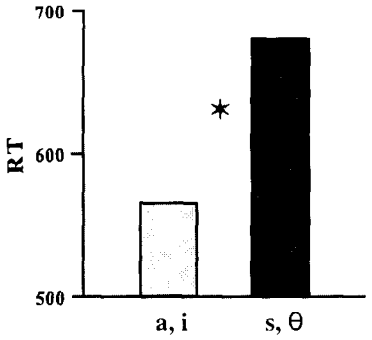

Fig. 10. Phoneme detection response times (msec) of Spanish listeners for vowel and fricative targets in Spanish words.

it fails to account for why fricatives are harder in Spanish. The explanation for that appears to lie in the relative information value of each type of phoneme. Consider for instance the way in which we spot a speaker's dialect. In English there are strong cues in the vowels (compare the Southern British, Northern British and Scottish pronunciations of *look*, *luck*, *luke*). Processing a vowel in English thus informs the listener not just about the identity of the phoneme and hence of the word of which it forms part, but also about the background of the speaker. In Spanish there is less to be learnt about the speaker's background from the vowels; but the fricatives can be much more informative (compare Castilian and South American pronunciations of *gracias*). In Spanish, in consequence, fricatives may call for more processing, and may for this reason be responded to more slowly in phoneme detection. In other words, by comparing results from several languages we again arrive at a universal conclusion: the difficulty of phoneme detection depends upon the information value of a given sort of phoneme in each individual language.

2.3. Phoneme assimilation

A third example, again from the area of phoneme perception, is provided by the phenomenon of regressive assimilation, whereby a phoneme adopts features of the immediately following phoneme. In *sunbathing*, for example, the /n/ may change into an /m/ under influence of the bilabial place of articulation of the following /b/, resulting in the pronunciation *sumbathing*. Assimilation rules of this sort are in some languages obligatory; in Japanese, for example, that is true of exactly this case, a nasal before a stop consonant (thus *kampai* could not possibly be pronounced *kanpai* or *kangpai*). In other languages – such as English or Dutch – the same rule can occur in optional form (thus it is hardly noticeable whether an English speaker says *sumbathing* or *sunbathing*).

A psycholinguist who wishes to discover whether assimilation affects the recognition of phonemes faces a problem of logic. The obvious questions are: does the presence of assimilation facilitate phoneme processing, and do violations of the assimilation rule have an adverse effect on processing. The first question can only be tested in a language in which the rule is optional, because if the rule is obligatory, the same word cannot be presented both with and without assimilation, for without is not an option. The second question, on the other hand, can only be tested in a language in which the rule is obligatory, because there is no such thing as a violation if the rule is optional. Thus the complete picture cannot be obtained from experiments in any one language alone, but only from experiments in more than one language. In our group we have recently looked at a number of cases of phoneme assimilation rules. First, we found that violations of the rule of homorganic assimilation of a nasal to a following stop consonant in Japanese do indeed adversely affect processing:

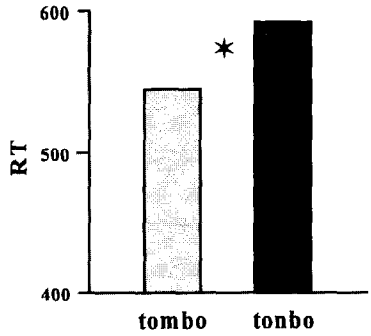

Fig. 11. Phoneme detection response times (msec) of Japanese listeners to post-nasal consonant targets in Japanese words with and without regressive assimilation of the nasal to the place of articulation of the following consonant (e.g. /b/ in *tombo* versus *tonbo*).

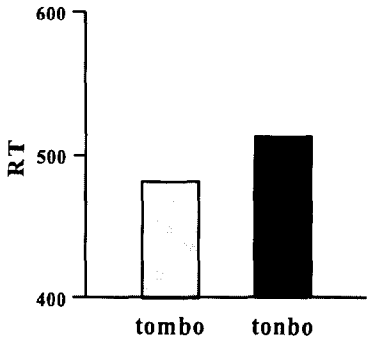

Fig. 12. Phoneme detection response times (msec) of Dutch listeners to post-nasal consonant targets in Japanese words with and without regressive assimilation of the preceding nasal to the place of articulation of the target consonant (e.g. /b/ in *tombo* versus *tonbo*).

the detection of a stop consonant such as /b/ is faster in *tombo* with assimilation than in *tonbo* without assimilation, i.e. with the rule violated (Fig. 11). If Dutch listeners are presented with the same material, however, they show no difference between the assimilated and the rule-violation case (Fig. 12), since for them, speakers of a language with the optional form of this rule, no violation is involved (Otake et al., 1996b).

Our next study concerned a second case of assimilation in Dutch, namely voice assimilation. In Dutch a sequence of obstruents does not undergo place assimilation (as is the case in English), but instead assimilates in voicing. Thus the Dutch word *kaas* (cheese), ending in /s/, may be pronounced *kaaz* in the word *kaasboer* (cheesemonger) because the following syllable 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 obligatory syllable-final devoicing rule. Our phoneme detection experiments (Kuijpers and Van Donselaar, Forthcoming) revealed no facilitatory effect of the assimilation: the /b/ in *kaasboer* was detected just as rapidly whether the word was pronounced *kaasboer* or *kaazboer* (Fig. 13). This is in accord with the result from the Japanese experiment with Dutch listeners: assimilation per se has no facilitatory effect. But a violation indeed adversely affected RTs: the /p/ in *kaasplank* (cheese board) was detected significantly more slowly when it was incorrectly pronounced *kaazplank* than when it was correctly pronounced *kaasplank* (Fig. 14). Again this is consistent with the results from the Japanese experiment in which violation of the obligatory assimilation rule caused slower responses from those listeners who commanded that rule, i.e. the Japanese listeners. In other words, a complete picture of the processing consequences of assimilation phenomena can be constructed as long as data are compared across languages with obligatory and optional assimilation.

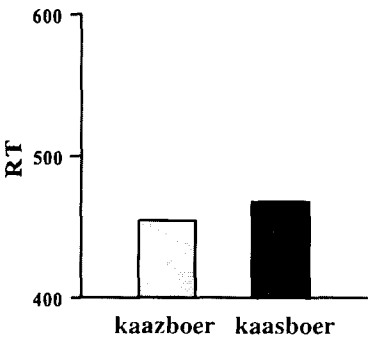

Fig. 13. Phoneme detection response times (msec) of Dutch listeners to consonant targets in Dutch words with and without regressive assimilation of the preceding obstruent to the voicing of the target consonant (e.g. /b/ in *kaazboer* versus *kaasboer*).

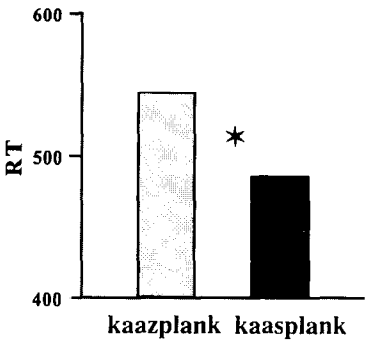

Fig. 14. Phoneme detection response times (msec) of Dutch listeners to consonant targets in Dutch words with and without inappropriate syllable-final voicing in a preceding obstruent (e.g. /p/ in *kaazplank* versus *kaasplank*).

3. Language-specific

The above three case studies have illustrated the first part of the argument outlined in the introduc-

tion: even universal features of linguistic structure can only be fully understood via comparison across languages. The second part maintains that even language-specific aspects of linguistic structure can prove extremely informative in the search for the universal model of language processing. Once again the argument will be illustrated with three case studies, each now from the area of spoken-word recognition. Instead of the simple detection tasks used in the experiments described in Section 2, the studies described below make use of a task which is particularly designed to study word recognition, in particular the recognition of words in continuous speech. This is word-spotting (Cutler and Norris, 1988), a psycholinguistic laboratory method in which listeners perform a task analogous to the word-spotting performed by automatic speech recognisers, namely location of real known words in a surrounding context. In the laboratory form of the task the context is kept to an absolute minimum: a syllable or even a single phoneme added to a real word. The subject in a word-spotting experiment is required to press the response button whenever a real known word is detected in the input. Thus the input might begin *crenthish*, *obzel*, *bookving* – and since the last item contains the real word *book*, subjects would press the button (and then also say the word *book* out aloud). What the subjects actually hear, in other words, is nonsense; but some nonsense items contain a word, and in that case a response is required. In the case of *bookving* the context *ving* follows the word; but it might also precede it, as in *kefapple* which consists of a context *kef* followed by *apple*.

In the past decade or so this task has provided many insights into the segmentation of continuous speech and the process of word recognition. For example, we have clear evidence that word recognition involves a process of competition between simultaneously activated word candidates (McQueen et al., 1994; Norris et al., 1995). Consider the two potential word-spotting items *domes* and *nemes*; both contain the embedded word *mess* preceded by a short context. They differ in that *domes* could be continued to form the English word *domestic*, while *nemes* cannot be continued to form a longer real word. In a word spotting experiment, listeners detected *mess* significantly more slowly in domes than in *nemes* (Fig. 15); in other words, the simultaneous

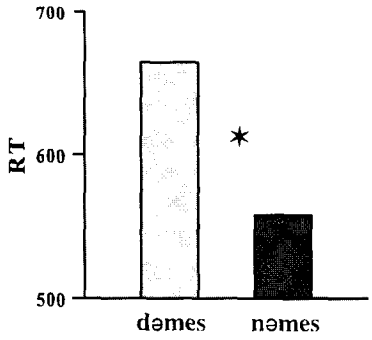

Fig. 15. Word spotting response times (msec) of English listeners to monosyllabic English words (e.g. *mess*) embedded in contexts which did or did not form the beginnings of other existing words (e.g. dəmes, nəmes).

activation of *domestic* as a potential word candidate interfered with the activation of *mess*. This interference effect is evidence for active competition between candidates for word recognition; not only are they activated by the input, they indeed compete in that the more words are activated, the less any individual word is initially activated. Exactly the same interference effect appeared in an analogous experiment in Dutch: the word *zee* (sea) is harder to spot in *muzee* than in *luzee* (Fig. 16), presumably because *muzee* can be continued to form *museum* whereas *luzee* cannot be continued to form any longer real word (Van Donselaar et al., Forthcoming a).

The vocabulary of any language is constructed from a very small number of phonetic building

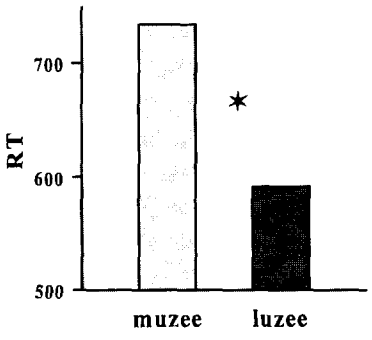

Fig. 16. Word spotting response times (msec) of Dutch listeners to monosyllabic Dutch words (e.g. *zee*) embedded in contexts which did or did not form the beginnings of other existing words (e.g. *muzee*, *luzee*).

blocks, in consequence of which all words have many phonetic neighbours which they closely resemble. The fact that speech occurs in the temporal dimension further entails that a recogniser is presented with the beginning of a word before the rest, and that beginning could also be the beginning of other words, unrelated to the actually uttered word (McQueen and Cutler, 1992; Cutler et al., 1994). The simultaneous activation of multiple potential words by the input leads to competition. It is reasonable to assume that such competition is part of the universal model of word recognition. Nevertheless, languages must differ in the type of initial input which activates lexical candidates, and indeed in the exact way in which the input produces lexical activation. These language-specific characteristics of lexical processing can nonetheless be extremely informative about the universal model, as the following three examples show.

3.1. Vowel harmony

The first example comes from Finnish, which exhibits strict constraints on the occurrence of vowels within words. If the first vowel in a Finnish word is /a/, /o/ or /u/, then no subsequent vowel in the same word may be /ae/, /oe/ or /y/; and vice versa. Thus *palo* is a word, as is *pöly*, but *paly* is impossible. It therefore follows that if two successive syllables contain vowels from these two mutually exclusive classes, the syllables cannot belong to the same word: there must be a word boundary between them. This vowel harmony phenomenon in Finnish therefore provides listeners with information which they could potentially exploit in solving the segmentation problem in continuous speech, that is, in finding where one word in a continuous speech signal ends and the next begins. The hypothesis that listeners could exploit vowel harmony mismatches in this manner was first proposed by Trubetskoy (1939); 56 years later it was put to experimental test by Kari Suomi, James McQueen and myself (Suomi et al., In press). In a word spotting experiment we presented Finnish subjects with words such as *palo*, with a minimal preceding context consisting of a single syllable in which the vowel either came from the same harmony class as the vowels of the embedded word, or from the other class. Thus *palo* occurred

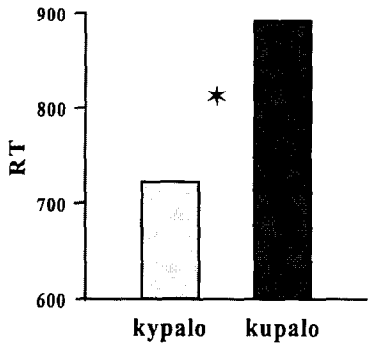

Fig. 17. Word spotting response times (msec) of Finnish listeners to Finnish words (e.g. *palo*) with preceding contexts containing harmonically mismatching versus matching vowels (e.g. *kypalo*, *kupalo*).

both in *kypalo*, with /y/ from the other class than the /a/ and /o/ of *palo*, and in *kupalo*, with /u/ from the same class. When there was a harmony class mismatch between the context and the embedded word, the subjects detected the word significantly more rapidly (Fig. 17) – the harmony mismatch, in other words, flagged the word boundary for them. Listeners thus seem to be able to exploit the language-specific effect of vowel harmony in segmenting speech, in very much the same way as language rhythm was exploited in the experiments described in Section 2.1.

Although vowel harmony of this kind is something which occurs in only a minority of the world's languages, nevertheless the results of this study are important in at least two ways for the universal model that we aim to construct. First, the kind of information which listeners are exploiting in this experiment is located not actually at the word boundary, but rather after it; it is the fact that the vowel in the second syllable belongs to a different class than the vowel in the first syllable which enables listeners to draw the conclusion that there must have been a word boundary preceding that second syllable. In this respect the exploitation of vowel harmony information resembles the exploitation of stress rhythm in segmentation by English listeners – in that case too the experiments showed that the vowel quality of a syllable led listeners to treat that syllable as the first syllable of a new word, in other words to conclude that a word boundary must have occurred immediately before it.

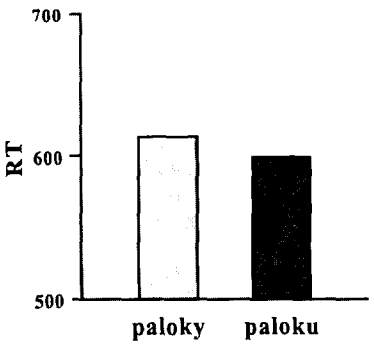

Fig. 18. Word spotting response times (msec) of Finnish listeners to Finnish words (e.g. *palo*) with following contexts containing harmonically mismatching versus matching vowels (e.g. *paloky*, *paloku*).

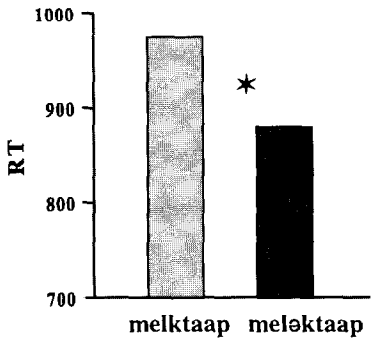

Fig. 19. Word spotting response times (msec) of Dutch listeners to Dutch words (e.g. *melk*) realised in underlying form or with optional epenthetic vowel.

Second, although harmony mismatches provided listeners with valuable information about the beginnings of words, the same mismatch at the end of a word did not prove of use. *Palo* was detected equally rapidly in either *paloky* or *paloku* (Fig. 18); listeners apparently did not need any assistance to find the end of a word, presumably because at that point the word was already fully activated. This finding is also consistent with findings in other languages, and the fact that the same pattern occurs in the exploitation of vowel harmony in Finnish and the exploitation of completely different sources of information in other languages suggests the involvement of a universal characteristic of the lexical access process.

3.2. Vowel epenthesis

The next example of a non-universal phenomenon which can potentially inform the universal model of language perception is vowel epenthesis, whereby in some languages certain consonant clusters are split by insertion of an intervening vowel. Examples from Dutch are [filəm] for *film*, [mɛlək] for *melk*. Because there is a universal tendency across languages towards alternation between vowels and consonants – such that in some languages such alternation is obligatory, and in no language anywhere is it illegal – splitting up consonant clusters with an epenthetic vowel in effect renders a language like Dutch somewhat more universal. It is therefore possible that there might be a processing advantage associated with epenthesis – if not for the speaker, who by virtue of the added phoneme in practice has to say more than is strictly necessary, then perhaps for the listener.

We investigated the effects of optional epenthesis on word recognition via a number of experiments (Van Donselaar et al., 1996, Forthcoming b; Kuijpers et al., 1996); these included a word-spotting study in which listeners were presented with words such as *melk*, with a minimal following context, and the embedded words themselves were either pronounced with epenthesis, thus [mɛlək], or without i.e. in the underlying form [mɛlk]. Subjects responded significantly faster with the word *melk* when they heard *meləktaap* than when they heard *melktaap* (Fig. 19). Thus although the words with the added epenthetic vowel were in fact longer than the same words in their underlying form, listeners recognised them more rapidly. (The visitor to Holland who fancies a glass of milk may thus obtain it more rapidly by asking for [mɛlək] rather than for [mɛlk]!)

3.3. Stress

The third and final example of language-specific phenomena is provided by stress – lexical stress, the prosodic structure of words such as observed in English. Across the world's languages, the prosodic structure of words is highly variable: in tone languages, lexical items may be distinguished by contrasts of pitch level or pitch contour on a single syllable; in pitch accent languages pitch contrasts are

drawn not between individual syllables but between polysyllabic words; in stress languages one syllable in a polysyllabic word is prosodically more salient, and this in turn may always occur in the same position within a word (in fixed stress languages) or may vary across words (in lexical stress languages such as English and Dutch).

Ten years ago an experiment in English showed that listeners made no use of purely prosodic information in initial lexical access. The evidence was that pairs of unrelated words distinguished only by prosody, such as *FORbear* and *forBEAR*, proved effectively homophonous; whichever of *FORbear* or *forBEAR* a listener heard, briefly both words were lexically activated. Thus it appeared that only the segmental information counted for lexical access; the purely prosodic differences which are all that distinguish *FORbear* from *forBEAR* did not contribute to the initial lexical activation stage. That result led the author of the article (Cutler, 1986) to the conclusion that "lexical prosody does not constrain lexical access". Does this conclusion however hold for all forms of lexical prosody?

Recent studies by Hsuan-Chih Chen and myself on Cantonese have shown that it is too strong a statement for tone languages, since our results (Cutler and Chen, 1995, 1997) suggest that tone contrasts are processed at the initial lexical activation stage. As described above, phonologically similar words that are simultaneously activated will actively compete with one another; we observed that similarities of tone and similarities of segmental structure contribute in just the same way to such competition effects. Thus the 1986 conclusion can obviously not be maintained for every kind of lexical prosody.

Lexical tone is realised on a single syllable, and tone contrasts can be exemplified by comparing two monosyllables. In this respect, tone differs from other forms of lexical prosody such as lexical pitch accent and lexical stress, which require polysyllabic strings in order for contrasts to be displayed. Thus it is conceivable that the domain of a prosodic phenomenon is of relevance for its role in lexical activation, i.e. that tone contrasts contribute to lexical activation because they can be fully realised on a single syllable, while stress contrasts fail to contribute because their domain is too large. This hypothesis can be tested by considering another form of lexical prosody with a polysyllabic domain: pitch accent in Japanese. Recent experiments by Takashi Otake and myself suggest that pitch accent behaves differently from stress. It is virtually impossible to determine the stress pattern of a Dutch or English word just from the initial syllable, at least if that syllable contains a full vowel (Jongenburger and Van Heuven, 1995). Japanese listeners, however, recognise pitch accent patterns with high accuracy from just the initial syllable of a polysyllabic word (Cutler and Otake, 1996). This suggests that the prerequisites for the use of pitch accent in word recognition in Japanese do exist, in contrast to the use of stress patterns in English.

If the claim that prosodic information is redundant in lexical activation holds not for all forms of prosody, and not even for all forms of lexical prosody which have a polysyllabic domain, then it would appear to hold only for stress patterns. Even that conclusion, however, turns out to be too strong. Consider the situation in which a word does not compete with phonologically similar words in the case that it is wrongly stressed; such a situation would imply that the word has not entered the competition process because it has not been activated, and if wrongly applied stress can prevent activation, then stress must be usable in lexical activation. In Dutch, indeed, this proves to be the case. The competition effect in Dutch described above, whereby the word *zee* is harder to detect in *muzee* than in *luzee* because *muzee* is the beginning of *museum*, disappears when that item is misstressed, i.e. pronounced *MUzee* (Fig. 20; Van Donselaar et al., Forthcoming a). Thus the fact that *MUzee* is stressed on the first instead of the second syllable is apparently sufficient to prevent activation of *museum* (with its stress on the second syllable) and hence to remove the competition of *museum* with *zee* which was seen with the input *muZEE* (cf. Fig. 16). This finding clearly implies that Dutch listeners can indeed make use of purely prosodic information in initial lexical activation.

The initial conclusion that lexical prosody does not constrain lexical access thus holds not for lexical prosody in general, not for prosody in a polysyllabic domain, and not even for stress per se, but only for stress in English. English stress is unusual in that pairs like *forbear* are extremely rare; stress contrasts

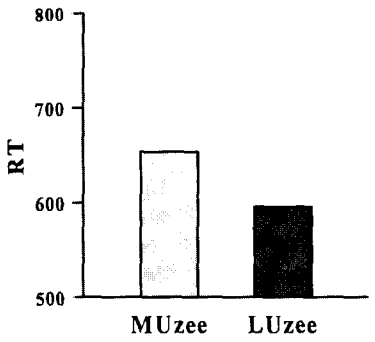

Fig. 20. Word spotting response times (msec) of Dutch listeners to monosyllabic Dutch words (e.g. *zee*) embedded in the same contexts as for Fig. 16, but with stress pattern mismatching the potential matrix word.

almost always exhibit segmental correlates, namely variation in vowel quality. Syllables are either stressed with a full vowel, or have a reduced vowel and are unstressed; unstressed syllables with a full vowel are extremely rare in English. In effect purely prosodic information is virtually redundant for ascertaining stress patterns in English.

This voyage through the different forms of lexical prosody has led finally to a clear conclusion which can inform the universal model of language processing, namely: prosodic information is of use in lexical activation precisely to the extent that it provides a unique and not otherwise replaceable contribution in reducing the number of competing potential candidate words. Prosodic information is not used in lexical activation in English simply because prosody in that language has so little to offer.

4. Conclusion

Stress, epenthesis and vowel harmony – these three language-specific (i.e. non-universal) aspects of linguistic structure have provided valuable information about universal aspects of linguistic processing. Three phonological effects which could in principle be investigated in any language, since they appear in all – syllables, vowels and consonants, assimilation phenomena – have by contrast only yielded their full processing story in the light of cross-linguistic comparison: the universal principle underlying the processing in question was only apparent once results from more than one language could be compared. Spoken-language processing by human listeners cannot, in effect, be understood unless it is studied comparatively. If everyone spoke the same language, perhaps life would be simpler in that language could no longer offer a source of conflict between peoples. It would perhaps be richer in that access to all the world's literary resources in the original language might encourage greater understanding of cultural differences. It might be easier in that scientists functioning in the international community would no longer be at an advantage or disadvantage, according to whether or not their native language was the dominant language of world science. But it would be poorer for those of us who work on human spoken-language processing: we would know far less about how the system works.

Acknowledgements

This is the text of a plenary lecture to the Fourth International Congress on Spoken Language Processing, Philadelphia, October 1996. The same lecture was presented (in Dutch) to the University of Nijmegen in September 1996, and published by Benda Drukkerij, Nijmegen, under the title "Eentaalpsychologie is geen taalpsychologie". Thanks are due to all the colleagues who collaborated on the work described here.

References

A. Cutler (1985), "Cross-language psycholinguistics", *Linguistics*, Vol. 23, pp. 659–667.

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

A. Cutler and H.-C. Chen (1995), "Phonological similarity effects in Cantonese word recognition", *Proc. Thirteenth Internat. Congress of Phonetic Sciences*, Stockholm, Vol. 1, pp. 106–109.

A. Cutler and H.-C. Chen (1997), "Lexical tone in Cantonese spoken-word processing", *Perception and Psychophysics*.

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

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

A. Cutler and T. Otake (1996), "The processing of word prosody in Japanese", *Proc. 6th Australian Internat. Conf. on Speech Science and Technology*, Adelaide, pp. 599–604.

A. Cutler, J. Mehler, D.G. Norris and J. Segui (1986), "The syllable's differing role in the segmentation of French and English", *J. Memory and Language*, Vol. 25, pp. 385–400.

A. Cutler, J.M. McQueen, H. Baayen and H. Drexler (1994), "Words within words in a real-speech corpus", *Proc. 5th Australian Internat. Conf. on Speech Science and Technology*, Perth, Vol. 1, pp. 285–288.

A. Cutler, B. Van Ooijen, D. Norris and R. Sánchez-Casas (1996), "Speeded detection of vowels: A cross-linguistic study", *Perception and Psychophysics*, Vol. 58, pp. 807–822.

W. Jongenburger and V.J. Van Heuven (1995), "The role of lexical stress in the recognition of spoken words: Prelexical or postlexical?", *Proc. Thirteenth Internat. Congress of Phonetic Sciences*, Stockholm, pp. 368–371.

C. Kuijpers and W. Van Donselaar (Forthcoming), "Effects of regressive assimilation on phoneme identification".

C. Kuijpers, W. Van Donselaar and A. Cutler (1996), "Phonological variation: Epenthesis and deletion of schwa in Dutch", *Proc. Fourth Internat. Conf. on Spoken Language Processing*, Philadelphia, PA, Vol. 1, pp. 149–152.

J.M. McQueen and A. Cutler (1992), "Words within words: Lexical statistics and lexical access", *Proc. Second Internat. Conf. on Spoken Language Processing*, Banff, Canada, Vol. 1, pp. 221–224.

J.M. McQueen, D.G. Norris and A. Cutler (1994), "Competition in spoken word recognition: Spotting words in other words", *J. Experimental Psychology: Learning, Memory and Cognition*, Vol. 20, pp. 621–638.

J. Mehler, J.-Y. Dommergues, U. Frauenfelder and J. Segui (1981), "The syllable's role in speech segmentation", *J. Verbal Learning and Verbal Behaviour*, Vol. 20, pp. 298–305.

J. Mehler, J. Bertoncini, E. Dupoux and C. Pallier (1996), "The role of suprasegmentals in speech perception and acquisition", in: T. Otake and A. Cutler, Eds., *Phonological Structure and Language Processing: Cross-Linguistic Studies* (Mouton/De Gruyter, Berlin), pp. 145–167.

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

T. Otake, G. Hatano, A. Cutler and J. Mehler (1993), "Mora or syllable? Speech segmentation in Japanese", *J. Memory and Language*, Vol. 32, pp. 358–378.

T. Otake, G. Hatano and K. Yoneyama (1996a), "Speech segmentation by Japanese listeners", in: T. Otake and A. Cutler, Eds., *Phonological Structure and Language Processing: Cross-Linguistic Studies* (Mouton/De Gruyter, Berlin).

T. Otake, K. Yoneyama, A. Cutler and A. van der Lugt (1996b), "The representation of Japanese moraic nasals", *J. Acoust. Soc. Amer.*

K. Suomi, J.M. McQueen and A. Cutler (In press), "Vowel harmony and speech segmentation in Finnish", *J. Memory and Language*.

N.S. Trubetskoy (1939), "Grundzüge der Phonologie", *Travaux du Cercle Linguistique de Prague*, Vol. 7.

W. Van Donselaar, C. Kuijpers and A. Cutler (1996), "How do Dutch listeners process words with epenthetic schwa?", *Proc. Fourth Internat. Conf. on Spoken Language Processing*, Philadelphia, PA, Vol. 1, pp. 94–97.

W. Van Donselaar, M. Koster and A. Cutler (Forthcoming a), "*Voornaam* is not a homophone: Lexical prosody and lexical access in Dutch".

W. Van Donselaar, C. Kuijpers and A. Cutler (Forthcoming b), "Vowel epenthesis in Dutch facilitates word processing".

B. Van Ooijen (1994), The processing of vowels and consonants, PhD Thesis, Leiden.

B. Van Ooijen, R. Sánchez-Casas, A. Cutler and D. Norris (Forthcoming), "Processing vowels and consonants in English and Spanish: The role of phonemic repertoire and variability".