SYLLABLE PROCESSING IN ENGLISH

Ruth Kearns, Dennis Norris* and Anne Cutler*[+]

* MRC Cognition and Brain Sciences Unit, Cambridge, United Kingdom
[+] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
anne.cutler@mpi.nl

ABSTRACT

We describe a reaction time study in which listeners detected word or nonword syllable targets (e.g. *zoo, trel*) in sequences consisting of the target plus a consonant or syllable residue (*trelsh, trelshek*). The pattern of responses differed from an earlier word-spotting study with the same material, in which words were always harder to find if only a consonant residue remained. The earlier results should thus not be viewed in terms of syllabic parsing, but in terms of a universal role for syllables in speech perception; words which are accidentally present in spoken input (e.g. *sell* in *self*) can be rejected when they leave a residue of the input which could not itself be a word.

1. INTRODUCTION

Syllables play an important role in language structure and in language use. In speech production, the syllable is the most efficient abstract unit on which to base phonological encoding [14]. In speech perception, syllables play a multitude of roles, some universal, some language-specific.

Questions of language-specificity concerning the syllable have prompted a very large body of research over the past few decades. Evidence from French [17] that the syllable functioned as a unit of segmentation, such that listeners process speech input syllable by syllable, could not be replicated in English [5]. This prompted the proposal that the asymmetry between the evidence from French and from English had it roots in rhythmic differences between the languages [7]; the syllable has a primary role in the rhythm of French, but English rhythmic structure is based on stress. On this proposal, segmentation would exploit the rhythmic structure of speech, and the syllable would thus be important for segmentation in languages in which it was also the basis of rhythmic structure. In languages with other rhythmic structures, those other structures would be important for segmentation.

Indeed, independent evidence had established a role for stress in the segmentation of English [1,8], and subsequent tests of the rhythmic hypothesis established a segmentation role for yet another language-specific rhythmic structure, namely moraic rhythm in Japanese [10,22]. In this line of research it is clear that where the syllable forms the basis of language rhythm, it accordingly plays a role in listening to speech; but because rhythmic structure differs across languages, the perceptual role of the syllable is thus also language-specific.

However, an important universal role for the syllable in spoken language understanding was proposed by Cutler, McQueen, Norris and Somejuan [4]; this proposal would thus apply to English as well as to other languages with different phonological structure. The universal role, they proposed, arises from the fact that the syllable is the smallest thing that can be a stand-alone word. This fact in turn constrains the operations of lexical activation and competition which form the basis of human speech recognition [11]. When human listeners hear speech, all word forms which are supported by the speech signal are activated and compete with one another for recognition; aligned and misaligned words compete (thus the input *least able* may activate *lease*, *stay* and *table* as well as the intended words) and recognition ensues when the actually spoken words triumph over their competitors. Norris, McQueen, Cutler and Butterfield [20] showed that spoken words are harder to recognise when abutted to a context consisting only of consonants (e.g. *egg* in *fegg*, *zoo* in *zooth*) than with a context consisting of a syllable (*egg* in *maffegg*, *zoo* in *zoothig*). They proposed that activation is subject to what they called a Possible Word Constraint (PWC), whereby activated candidate words are penalised if the effect of accepting them would be to leave a residue of the speech signal which was not itself viable as a candidate word. Because consonants cannot be stand-alone words, the consonant context in *fegg* and *zooth* fails this viability constraint, so that the activation of *egg* and *zoo* is reduced and the words are harder to recognise. The contexts *maff* and *thig*, in comparison, also happen not to be actual words, but might have been English words and therefore do not trigger any penalty on the words abutting them.

This viability constraint would have the effect of reducing unwanted activation of words embedded in other words - for instance, *east* in *least* would be penalised because /l/ could not be a word by itself. Only syllable contexts pass the constraint. Any syllable will do; importantly, it is not necessary that a syllabic context consist of a viable stand-alone syllable in the particular vocabulary under study. In English, open syllables with lax vowels are not viable stand-alone syllables (in contrast to many other languages, e.g. French, which has words such as *pas*, *thé*). But such syllables do not trigger a penalty on adjacent words for English listeners [21, see also 2]. Thus the PWC appears not to be subject to constraints of language-specific lexica, but to be potentially universal; hence, the syllable has a universal role in language processing.

However, proponents of a universal segmentation role for the syllable may argue that a segmentation account also predicts the experimental results on which the claims for the PWC are based. If listeners perforce process input syllable by syllable, would not *fegg* and *zooth* be processed as unitary wholes, but *maffegg* and *zoothig* be processed in two chunks: *maff-egg*, *zoo-thig*? Would not the words with consonant context be hard to recognise because the syllabic unit would have to be further decomposed, while the words with syllabic context would become automatically available as soon as the syllabic parsing mechanism separated them from their contexts?

If this account were to hold, then the asymmetry observed by Norris et al. [20] should appear also in a task involving explicit syllable segmentation. Norris et al.'s results were obtained with word-spotting [15], in which the task is to spot any real word embedded in heard nonwords; this task specifically addresses the recognition of real words in context. Syllable segmentation is usually studied with the fragment detection task [12], in which listeners respond when they detect a pre-specified target sequence, usually in initial position within a word or nonword. In the present study we test the syllabic segmentation explanation for the results of Norris et al. [20] by presenting the materials of that study to listeners instructed to perform fragment detection.

2. FRAGMENT DETECTION STUDY

2.1. Materials and Procedure

The stimuli used in this experiment were a subset of the materials used by Norris et al. [20]. The critical stimuli in that study were based on 48 monosyllabic and 48 bisyllabic words. Each word was abutted to a context such that a nonword resulted, consisting of the word plus a syllabic or consonantal residue. For half of the words the residue preceded the word (e.g. *fegg*, *maffegg*), for half it followed (e.g. *zooth*, *zoothig*).

The current experiment used only the 24 monosyllabic words with following residues. These could begin with a single consonant or a cluster, and could end with a vowel or a single consonant (e.g. *shoe, plough, run, spell*). An additional set of 24 monosyllabic nonword stimuli was generated from the non-target filler stimuli in that study (e.g. *voo, pel, prul*). The nonwords were also followed by either a syllable or a consonant residue. Nine words and 13 nonwords had a C(C)V structure, and 15 words and 11 nonwords were C(C)VC.

Word and nonword targets were each presented four times to each subject. On two trials the target was paired with one of its own nonword matrices (BELL-*belshig*), and on the remaining two it was paired with the matrices from a matched word or nonword target (BELL-*trelshek*). Each target therefore appeared twice in a YES (target present) trial and twice in a NO (target absent) trial. Within these, it appeared once with a syllable residue and once with a consonant residue. Thus each subject heard each target-matrix combination once. In the NO trials, target specifications and embedded targets were paired such that word and nonword target specifications were equally likely to precede a word or nonword embedded target. In 12.5% of the NO trials the target partially overlapped the beginning of the nonword (e.g. *zee-zooth*; *fun-faudgul*); such trials discourage anticipatory responding [19]. Examples of the stimuli are shown in Table 1. There were 192 experimental items in total (96 YES, 96 NO).

Table 1. An example word-nonword stimulus pair and its arrangement across the two stimulus blocks.

Block A			Block B	
BELL	bellsh		BELL	belshig
BELL	trelsh		BELL	trelshek
TREL	bellshig		TREL	belsh
TREL	trelshek		TREL	trelshek

Stimuli were divided into two blocks. Half of the subjects heard the blocks in one order, and half in the other order. For each of those two groups of subjects there were two different randomised presentations of the items within the blocks. Each block contained an equal number of items with phoneme or syllable residues, and an equal number of word and nonword, and YES and NO targets. In each block each target appeared with either only a syllable residue or only a phoneme residue. There were 6 practice trials, and two filler trials between the end of the practice and the start of the experimental items.

All stimuli were recorded onto DAT tape using a high quality microphone in a sound attenuated booth. Target specifications (syllables) were spoken by a female speaker of British English and the nonwords were spoken by a male speaker of British English. (The latter recording was in fact that used in Norris et al.'s [20] Experiment 1.) The stimuli were digitally transferred to a computer where they were downsampled to 22.05kHz and edited before being transferred to compact disc.

Each trial consisted of the presentation of a 100ms 1kHz tone, followed 1000ms later by the target specification. The nonword started 1000ms after the onset of the target specification. There was an interval of about 2500ms between trials. Subjects were instructed to press a button as quickly as possible if the nonword started with the specified target. Stimuli were presented over headphones from a portable CD player. Response timing was controlled by the TSCOP software [18] on a portable computer. Responses were measured from the onset of the nonword.

2.2. Subjects

Subjects were 24 undergraduates from Jesus College, Cambridge, who were each paid a small honorarium for participating. All were native speakers of British English without reported hearing deficits.

2.3. Results

Table 2 presents overall mean RTs and errors (in parentheses), as a function of lexical status of target (word, nonword) and type of residue (syllable, consonant). The overall false alarm rate on trials where the target was not present was 2.8%. Analyses of variance were conducted on both RTs and error rates, separately across subjects and across items.

Table 2. Mean fragment detection RT (ms) and percent errors, as a function of lexical status of target (word, nonword) and type of residue (syllable, consonant).

	Syllable	Consonant
Word	472 (0.6)	477 (2.7)
Nonword	463 (1.5)	464 (2.9)

In the RT analysis there was no effect of residue (both $Fs < 1$), nor of lexical status ($F1(1,19) = 1.44$, $p > .2$; $F2 < 1$). In the error analysis the main effect of residue was significant in the subjects analysis, but not the items analysis ($F1(1,19)=6.77$, $p<.02$ $F2(1,92)=2.44$, $p>.1$).

The target items in this experiment included both CV and CVC syllables. Fragment detection experiments that have directly compared monitoring for CV and CVC targets have generally found different patterns of reaction times to the two

kinds of target [6]. We therefore performed a further analysis to examine CV and CVC targets separately. Table 3 shows the RTs and error rates, broken down by target structure. As can be seen from the table, the two types of target differed in the effect of residue type. CV targets were detected more accurately and 13 ms more rapidly with consonant residues than with syllable residues, while CVC targets were responded to much more accurately and 16 ms faster with syllable residues than with consonant residues. (This pattern was the same for word and for nonword targets.)

Table 3. Mean fragment detection RT (ms) and percent errors, as a function of target structure (CV,CVC) and type of residue (syllable, consonant).

	Syllable	Consonant
CV	473 (1.2)	460 (0.9)
CVC	464 (0.8)	480 (4.5)

In the RT analysis the interaction between residue and target structure was significant in the subjects analysis only ($F1(1,19)=6.39$, $p< .025$; $F2(1,44)=2.85$, $p< .1$). In the error analysis the interaction was significant by subjects, and marginally significant by items ($F1(1,19)=7.22$, $p< .02$; $F2(1,44)=3.92$, $p< .055$).

3. COMPARISON OF FRAGMENT DETECTION AND WORD-SPOTTING

In each of the word-spotting experiments of Norris et al. [20], the same real-word materials used in the present study were detected substantially more slowly and less accurately with the consonantal than the syllabic residues; the smallest differences they observed for these materials in any experiment were 71 ms of RT and 7% errors. That is clearly a very different result from the overall 5 ms RT difference and 2.1% error difference observed here with fragment detection. No analysis of the word-spotting data had been undertaken as a function of phonological structure of the embedded word; we carried out such an analysis of Norris et al.'s Experiment 1, for comparison with the present data. The results are displayed in Table 4.

Table 4. Mean word-spotting RT (ms) and percent errors, from Norris et al. [20], as a function of word structure (CV,CVC) and type of residue (syllable, consonant).

	Syllable	Consonant
CV	971 (28.3)	1056 (39.4)
CVC	854 (28.6)	990 (38.6)

As can be seen, the CV (*zoo*) and CVC (*bell*) words do not differ in word-spotting: both show a very large disadvantage (more than 80 ms and 10% of error) for consonant residues. The interaction of these two factors is insignificant ($F<1$).

A further analysis examined, item by item, the correlation of the differences between the syllable and consonant conditions in the word-spotting study and in the present fragment detection study. This too produced an insignificant result ($r = 0.19$); that is, the results of the two studies are not closely correlated.

4. LEXICAL STATISTICS

The rationale that Norris et al. [20] proposed for the PWC is the reduction of spurious competition in spoken language understanding; recognition of intended words can potentially be speeded if word forms which are accidentally present in the speech signal, because they are fully embedded in or across the intended words, have their competitive power reduced. In this context it is relevant to ask how much embedding occurs in the vocabulary. Four out of five English polysyllabic words contain embedded words; 50% of polysyllabic words begin with a fully embedded other word [16]. The analysis from which these results are taken considered only embeddings in which syllable boundaries of matrix and embedded word were aligned (e.g. *can* in *canvas*, or *apart* in *apartment*). Analyses of real-speech corpora [3] however show large differences between embedding with vs. without syllable boundaries being maintained. The proportion of embedding without respect to syllable boundaries (e.g. *can* in *cant*) rose to over 90% of all words.

Thus the PWC really does help remove spurious competition. For the present study we further calculated how many vowel-final (e.g. *zoo*) and consonant-final (e.g. *bell*) monosyllables in English occur as initial embeddings with only subsequent consonant(s) (e.g. *zoom*, *belch*) versus with subsequent syllable(s) (*zulu*, *bellows*).

The results are shown in Table 5. Although there are more word types embedded with syllable boundaries aligned (e.g. *zoo* in *zulu*, *bell* in *bellows*), the frequency-weighted CELEX token counts tell a different story. In agreement with the earlier results from speech corpora analyses [3], these show that words embedded with only consonantal residue (*zoo* in *zoom*, *bell* in *belch*) constitute overall more than 64% of the cases. In other words, application of the PWC in recognition would remove at a stroke almost two-thirds of all spurious embeddings of short words in word-initial position (the most common type of embedding [3,16]).

Table 5. Number of types and (in parentheses) tokens per 42.4 million words of existing English vowel-final (CV) and consonant-final (CVC) monosyllabic words embedded at the onset of other words, as a function of type of residue.

Residue	Syllable(s) e.g. *zoo* in *zulu*	Consonant(s) e.g. *zoo* in *zoom*
CV	13360 (64373)	9175 (106303)
CVC	16463 (21519)	8922 (49037)

5. DISCUSSION

Our findings give no support to the notion that English listeners segment speech input syllable by syllable. Overall, no advantage appeared in the present results for target fragments which were aligned with syllable boundaries over those which were not. The results are thus consistent with previous counter-evidence for syllabic segmentation in English [5], and in particular counter the argument that word-spotting differences observed with the very same recordings used here could have arisen from syllabic segmentation.

No difference appeared in our results for real-word versus nonword targets. However, we did observe an interesting difference between vowel-final (CV) versus consonant-final (CVC) targets - a difference which did not appear in the original

word-spotting study. For CV targets syllable residues made detection harder, while for CVC targets consonant residues made detection harder. We believe that simple phonetic explanations apply to each of these findings separately. Each concerns detection of the final phoneme of the target - note that syllable detection can only occur once the entire target has been successfully processed, so that final phoneme effects will be particularly relevant to the determination of RT.

Vowels in monosyllables tend to be longer than vowels in the first syllable of bisyllables [13], and longer vowels are responded to more rapidly in detection tasks than shorter ones [9]. Thus the vowel portion of *zoo* would be more rapidly detected and the response more rapidly issued in *zooth* than in *zoothig*. While this vowel difference would also hold for CVC targets, the vowel is in this case not the final phoneme of the target. Instead, the final phoneme is a consonant which in the case with a consonant residue (*bell* in *belsh*) forms a coda cluster with the residue. We suggest that this additional step of decomposing a coda cluster in order to identify the final phoneme of the target is responsible for the difficulty of consonant residues for CVC targets.

This potentially constitutes evidence for a processing role for within-syllable structure, as other researchers have claimed [23]. However, we reiterate that our principal finding is that an explicit syllabic segmentation task produces a radically different pattern of results than is observed with word-spotting given identical input. The role of the syllable itself in English is thus not as a segmentation unit. Instead, we view the syllable in the terms proposed by Norris et al. [20], namely as a viability filter in the lexical activation process. This role is apparently played universally by the syllable, since syllables which can or could not stand alone as words of the language under test are equally acceptable as word-spotting residues [2,21]. We propose that this function for the syllable has its roots in the fact that universally, the syllable is the smallest unit which can stand alone as a word. The speech recognition process can benefit from this knowledge by ruling out a priori activated word candidates which would leave lexically unviable residues.

6. ACKNOWLEDGEMENTS

We thank Sally Butterfield and James McQueen for extensive assistance with this project. Ruth Kearns is now with Procter and Gamble International Operations SA.

7. REFERENCES

[1] Cutler, A. and Butterfield, S., "Rhythmic cues to speech segmentation: Evidence from juncture misperception", *J. Mem. and Lang*., 31: 218-236, 1992.

[2] Cutler, A., Demuth, K. and McQueen, J.M., "Universality versus language-specificity in listening to running speech", *Psych. Sci.,* 13: 258-262, 2002.

[3] Cutler, A., McQueen, J., Baayen, H. and Drexler, H., "Words within words in a real-speech corpus". *Proc. of the 5th Australian International Conference on Speech Science and Technology*, Perth, Vol. 1, 362-367, 1994.

[4] Cutler, A., McQueen, J.M., Norris, D. and Somejuan, A., "The roll of the silly ball", in: *Language, Brain and Cognitive Development: Essays in honor of Jacques Mehler*, ed. E. Dupoux (pp. 181-194), MIT Press, Cambridge, MA, 2001.

[5] Cutler, A., Mehler, J., Norris, D. and Segui, J., "The syllable's differing role in the segmentation of French and English", *J. Mem. and Lang*., 25: 385-400, 1986.

[6] Cutler, A., Mehler, J., Norris, D. and Segui, J., "Phoneme identification and the lexicon", *Cogn. Psych*., 19: 141-177, 1987.

[7] Cutler, A., Mehler, J., Norris, D. and Segui, J., "The monolingual nature of speech segmentation by bilinguals", *Cogn. Psych*., 24: 381-410, 1992.

[8] Cutler, A. and Norris, D., "The role of strong syllables in segmentation for lexical access", *J. Exp. Psych.: Human Percept. and Perf.,* 14: 113-121, 1988.

[9] Cutler, A., Ooijen, B. van, Norris, D. and Sánchez-Casas, R., "Speeded detection of vowels: A cross-linguistic study", *Percept. & Psychophysics,* 58: 807-822, 1996.

[10] Cutler, A. and Otake, T., "Mora or phoneme? Further evidence for language-specific listening", J. *Mem. and Lang*., 33: 824-844, 1994.

[11] Frauenfelder, U.H. and Floccia, C., "The recognition of spoken words", in: *Language Comprehension, A Biological Perspective*, ed. A. Friederici (pp. 1-40), Springer, Heidelberg, 1998.

[12] Frauenfelder, U.H. and Kearns, R.K., "Sequence monitoring", *Lang. and Cogn. Proc*., 11: 665-673, 1996.

[13] Lehiste, I., "The timing of utterances and linguistic boundaries", *J. Acoust. Soc. of Am*., 51: 2018-2024, 1972.

[14] Levelt, W.J.M., Roelofs, A. and Meyer, A.S., "A theory of lexical access in speech production", *Behav. and Brain Sci.,* 22: 1-38, 1999.

[15] McQueen, J., "Word spotting", *Lang. and Cogn. Proc*., 11: 695-699, 1996.

[16] McQueen, J.M., Cutler, A., Briscoe, T. and Norris, D., "Models of continuous speech recognition and the contents of the vocabulary", *Lang. and Cogn. Proc*., 10: 309-331, 1995.

[17] Mehler, J., Dommergues, J.-Y., Frauenfelder, U. and Segui, J., "The syllable's role in speech segmentation", *J. Verbal Learning & Verbal Behav*., 20: 298-305, 1981.

[18] Norris, D., "A computer-based programmable tachistoscope for non-programmers", *Behav. Res. Methods, Instr. and Comp*., 16: 25-27, 1984.

[19] Norris, D. and Cutler, A., "The relative accessibility of phonemes and syllables", *Percept. & Psychophysics,* 43: 541-550, 1988.

[20] Norris, D., McQueen, J.M., Cutler, A. and Butterfield, S., "The possible-word constraint in the segmentation of continuous speech", *Cogn. Psych*., 34: 191-243, 1997.

[21] Norris, D., McQueen, J.M., Cutler, A., Butterfield, S. and Kearns, R., "Language-universal constraints on speech segmentation", *Lang. and Cogn. Proc*., 16: 637-660, 2001.

[22] Otake, T., Hatano, G., Cutler, A. and Mehler, J., "Mora or syllable? Speech segmentation in Japanese", *J. Mem. and Lang*., 32: 358-378, 1993.

[23] Treiman, R., "On the status of final consonant clusters in English syllables", *J. Verbal Learning and Verbal Behav*., 23: 343-356, 1984.