COGNITIVE PSYCHOLOGY **33,** 111–153 (1997)
ARTICLE NO. CG970649

Bootstrapping Word Boundaries: A Bottom-up Corpus-Based Approach to Speech Segmentation

Paul Cairns and Richard Shillcock

Centre for Cognitive Science, University of Edinburgh, Edinburgh, United Kingdom

Nick Chater

Department of Psychology, University of Warwick, Coventry, United Kingdom

and

Joe Levy

Department of Psychology, Birkbeck College, University of London, London, United Kingdom

Speech is continuous, and isolating meaningful chunks for lexical access is a nontrivial problem. In this paper we use neural network models and more conventional statistics to study the use of sequential phonological probabilities in the segmentation of an idealized phonological transcription of the London–Lund Corpus; these speech data are representative of genuine conversational English. We demonstrate, first, that the distribution of phonetic segments in English is an important cue to segmentation, and, second, that the distributional information is such that it might allow the infant, beginning with only a sensitivity to the statistics of subsegmental primitives, to bootstrap into a series of increasingly sophisticated segmentation competences, ending with an adult competence. We discuss the relation between the behavior of the models and existing psycholinguistic studies of speech segmentation. In particular, we confirm the utility of the Metrical Segmentation Strategy (Cutler & Norris, 1988) and demonstrate a route by which this utility might be recognized by the infant, without requiring the prior specification of categories like ''syllable'' or ''strong syllable.''

INTRODUCTION

One of the first problems the infant must resolve in developing a linguistic competence is the speech segmentation problem: the continuous speech stream

This work was supported by ESRC (UK) Grants R000 23 3649 and R000 22 1435. We thank John Morton, Dennis Norris, and Anne Christophe for constructive comments on earlier drafts of this paper; all remaining errors are our own. Address correspondence and reprint requests to Dr. Richard Shillcock at the Centre for Cognitive Science, 2 Buccleuch Place, Edinburgh EH8 9LW, United Kingdom. Telephone: +441 31 650 4425. E-mail: rcs@cogsci.ed.ac.uk.

0010-0285/97 $25.00

must be analyzed into its constituent words and morphemes. In this paper we are concerned with the psychological questions of the nature of the processor that the infant brings to the problem of segmentation and the way in which it develops an eventually adult capacity to segment speech. Does the infant require, at the outset of developing a speech processing capacity, an inbuilt sensitivity to certain structures, such as syllables or words, for instance, or to rhythmic structure itself? Or can the relevant linguistic categories be discovered in the course of exposure to continuous speech? What linguistic units are important to the development of segmentation, and how large a window onto the speech is useful? We address these issues principally by means of a comprehensive statistical analysis of a large, representative corpus of transcribed English speech. Such an analysis reveals more about the nature of the segmentation problem and, by implication, the nature of the processor necessary to solve it.

We will argue that the infant initially requires only a general sensitivity to the statistics of its environment for the structural regularities of English speech to begin to emerge and be recognized; the infant then requires the ability to instantiate these regularities in terms of new levels of representation and processing. We will take the starting point to be a subsegmental representation of the speech stream and the goal to be that same speech stream correctly divided into words. Different statistical approaches are appropriate to subsegmental and segmental levels of representation, so we employ below a connectionist modeling approach to the former and more conventional statistical approaches to the latter. The success of any one approach in segmenting the speech data represents an existence proof that the problem is tractable by a processor that makes only the same assumptions as the successful approach.

THE SEGMENTATION PROBLEM

Contrary to the impression that we have when listening to speech, most spoken words are not clearly delineated by acoustic gaps. Segmentation is the process by which the listener divides up the continuous speech stream into linguistically and psychologically significant units that can be used to access meaning. As in other domains such as vision (e.g., Bhattacharjya & Roysam, 1994) and audition (e.g., Doutriaux & Zipser, 1991), where signals must be segmented into meaningful chunks, segmentation and recognition appear to stand in a chicken-and-egg relationship. The identification of a particular stretch of speech as a meaningful unit presupposes recognizing what that unit is, but recognition seems to be possible only once segmentation has been carried out.

There are two ways in which to break out of the segmentation–recognition circle. The first is to let the system put forward tentative hypotheses concerning segmentation and recognition and reinforce those hypotheses that fit together. The system may therefore iteratively settle on an appropriate segmen-

tation of the input and at the same time recognize the units of interest: the strategy is thus synergistic. The second approach is to attempt to find reliable cues to segmentation, which are independent of the identity of what is being segmented. According to this approach, segmentation can be carried out bottom-up and its output fed on to later recognition processes. Researchers have explored both of these approaches in the context of psycholinguistic theory and speech technology.

The two approaches exemplify a fundamental issue within cognitive psychology that has been much debated in recent years: whether parts of the human language processor are interactive or modular (see, e.g., Elman & McClelland, 1988; Fodor, 1983; Massaro, 1994; Tanenhaus & Lucas, 1987). The fundamental difference between interactive and modular accounts is that only the former allow top-down influence of higher-level information on lower-level processing. However, it is important to note that both accounts can make use of low-level information, and thus a successful bottom-up computational modeling attempt cannot directly refute the top-down account, in favor of the modular account, although it may be argued that the encapsulated bottom-up explanation is the more parsimonious explanation.

There is a second segmentation problem: how can an infant learn a segmentation strategy in the first instance? Whatever the merits of interactive and bottom-up accounts of adult speech segmentation mechanisms, in development it seems that purely bottom-up cues must be used, at least initially, since the child has not yet learned the linguistic units upon which interactive models rely (Mehler, Dupoux, & Segui, 1990). These units may differ among languages such as English, French, and Japanese. Even if one posits an adult interactive strategy, this will be practicable only once a lexicon of a certain critical size is acquired by data-driven means. Although one might propose a set of genetically defined parameters which permit selection of a language-specific segmentation strategy, the infant can apply such innate principles only once some preliminary analysis of the raw speech stream has been developed. Even strongly innatist theories make use of ''triggers'' that are activated by input structures; knowledge of these structures must be bootstrapped (see, e.g., Lightfoot, 1991). We argue below that algorithmically simple distributional mechanisms are important early in language development, and, in particular, in learning to segment the speech stream. The simplicity of these mechanisms, and of their constituent representations, is a reason for them to be preferred over rival mechanisms. It remains a possibility that some aspects of this early distributional analysis are innately specified; in the modeling we describe below, the subsegmental representations, for instance, are possible candidates for such an innate specification.

Until recently, the importance of distributional information in acquisition has been eclipsed by the focus on innate constraints on language learning and formal, structural analysis of language. Now, however, the study of simple

distributional mechanisms is a focus of research in computational linguistics (Charniak, 1993) and psychology (Elman, 1990; Finch & Chater, 1993). In this paper we explore a possible source of bottom-up information for bootstrapping lexical boundary detection in English, which might be used developmentally and which relies on exploiting the distributional statistics of phonetic segments. We consider whether constraints on word boundaries can be inferred from a continuous stream of segments. We suggest that distributional statistics are a useful source of constraint, which in conjunction with other information allow the infant to learn to segment successfully. We also consider whether similar mechanisms can play a role in adult segmentation. Credence can be given to the bottom-up account if we can demonstrate its computational viability using genuine data. Indeed, a statistical model of learning can make claims about human behavior only to the extent that its input is representative of real spoken language; to this end we use a large corpus of transcribed speech as input for our models. We do not claim that distributional information alone is involved in either development or adult performance, although we suggest that distributional information must play a major role in the earliest stages of acquisition. Our methods are sufficiently simple to be realized by general cognitive learning mechanisms. In other languages, of course, other information may be important in beginning the process of segmentation.

The structure of this paper is as follows: first, we describe models that exemplify the two main approaches to breaking out of the segmentation–recognition circle that have been put forward in the literature. Not all of these models have been intended as models of development, but it is desirable to explore the extent to which they lend themselves to developmental explanation, before looking for new theories. We consider closely the metrical strategy put forward by Cutler, Norris, and colleagues (e.g., Cutler & Norris, 1988). Second, we present both supervised and unsupervised statistical *n*-gram models for segmentation, based on an input composed of phonological segments. Third, we consider an input based on subsegmental primitives, and we present a neural network for modeling segmentation. Finally, we discuss the models and their implications for human behavior.

INTERACTIONIST ACCOUNTS

The interactionist position can be traced to the original version of the Cohort Model of Marslen-Wilson (Marslen-Wilson & Welsh, 1978) and to other strictly sequential models of word recognition (e.g., Cole & Jakimik, 1980). According to this account, it is the lexeme that licenses segmentation. As the speech input arrives over time, the words that become incompatible with the input are incrementally eliminated until one winning candidate emerges. The stored lexical phonology of the winner specifies the boundary at which the next word begins. Thus, after the initial portion of the input

has activated the lexical cohort, the information that facilitates segmentation crucially flows from the lexical level and therefore is essentially top-down. The Cohort Model formed the basis of the connectionist TRACE model of McClelland and Elman (1986), in which activation from competing lexical nodes serves to cut up the input; lexical segmentation is a by-product of word recognition in this view (see Frauenfelder & Peeters, 1990).

As stated above, it is difficult to reconcile the interactionist approach with the development of segmentation, since a lexicon is presupposed. However, Suomi (1993) has suggested a developmental model in which recognition occurs in a Cohort-like manner. The segmentation problem is avoided by proposing that the child at first learns words that are spoken in isolation and that as each word is added to the lexicon it becomes available for use in helping to segment future input interactively. However, there is no formal demonstration of the viability of this approach; it is unclear how many items would have to be acquired in isolation before successful interactive segmentation could predominate. Some evidence is required that a sufficient number of items are actually spoken as isolated words, or are isolated by pausing, in speech to infants: in the CHILDES corpus (Macwhinney & Snow, 1985) around 15% of utterances that children hear are single words. Tokens of words spoken in isolation tend to be phonologically less reduced and cannot be subject to assimilation across word boundaries, compared with tokens that are embedded in phrases (see Jusczyk, 1993a), meaning that the former may not necessarily make accurate templates for the recognition of the latter. Because of these potential problems, it is both desirable and likely that the infant has some means of bootstrapping the recognition of lexical junctures.

BOTTOM-UP ACCOUNTS

Bottom-up accounts seek to segment on the basis of information in the speech signal and preclude any top-down contextual influence from the lexicon. The problematic nature of bottom-up segmentation became amply evident when speech technologists attempted to create automatic bottom-up procedures to extract words from spoken input (e.g., Cole, 1980): it was found that a broad conspiracy of various levels of information contributes to segmentation, from acoustic to discoursal. The fact that in casual fluent speech, the majority of word boundaries is not clearly delimited by acoustic discontinuities impedes attempts at automatic segmentation (see Klatt, 1980).

There have been three main approaches to bottom-up segmentation, making use of different cues. The first approach has addressed phonetic juncture marking, such as aspiration, or alterations in voice onset timing (e.g., Church, 1987; Lehiste, 1971). Such phonetic detail is not considered further in this paper. The second approach has been concerned with prosodic marking that specifies the initial portion of a word, given an already syllabified input (e.g., Cutler & Norris, 1988; Cutler, 1993; Cutler & Butterfield, 1992; Grosjean &

Gee, 1987). The third approach has employed distributional information, the use of sound sequence probabilities to predict likely junctures (e.g., Church, 1987; Harrington, Watson, & Cooper, 1988). The second and third approaches are discussed in detail below.

Whereas many models use only one of these types of cue, there is no reason why various types of information should not be integrated; indeed, this is what we believe to be necessary for a complete account of segmentation. Although at present we restrict our investigation to the use of distributional information, we do not claim that this is the only sort of information that can be brought to bear in solving the segmentation problem.

In discussing models that purport to be bottom-up, it will be useful to make a distinction between what we shall call ''weakly bottom-up'' and ''strongly bottom-up'' systems. A weakly bottom-up system is one in which higher-level information is not used during processing, although it can be used to teach the system or can be implicit in its construction. As an example, Norris's connectionist model of spoken word recognition (Norris, 1990, 1992, 1993) would be classified as weakly bottom-up: when the model is trained, lexical information is employed in order to back-propagate error, so each connection in the network will come to encode at least some lexical knowledge. Therefore, even though the network has no connections from lexical to phoneme nodes, lexical knowledge is still employed when it recognizes words, only this knowledge is learned and implicit, as opposed to imposed and explicit. A strongly bottom-up system, on the other hand, is one in which higher-level information is not used, either in training or in operation. Only a strongly bottom-up system can potentially provide a good model of the development of segmentation if we assume an empiricist position in which higher-level information must be bootstrapped from primary, low-level data.

The Prosodic Marking Approach

An approach that seems to be strongly bottom-up is that of Cutler, Norris, and colleagues (Cutler, Mehler, Norris, & Segui, 1986; Cutler & Norris, 1988; Cutler, 1986, 1993). The Metrical Segmentation Strategy (MSS) holds that listeners tend to segment input before strong syllables in English. More precisely, when a full vowel is heard, a boundary is hypothesized at the beginning of the syllable of which the vowel is nucleus, and it is that syllable that the listener uses in an attempt at lexical access. Schwa (/ə/) always produces a weak syllable, but other vowels may be realized with reduced forms, too. Thus the listener segments speech by making reference to an acoustic/phonetic cue, but this cue indicates only the general location of the boundary rather than the exact juncture point as is the case with the cues investigated by Lehiste (1971) and others. Note that strong and weak syllables are not solely specified by lexical stress marking. To illustrate this, Cutler (1993) gives as examples: *generous, generic, generate, generation.* These items have the

following metrical strong–weak patterns: SWW, WSW, SWS, and SWSW, respectively.

Cutler and Carter (1987) have provided evidence for the potential of the MSS by examining the statistics of a lexicon and an orthographically transcribed corpus of speech. They show that around 83–90% of open-class word types have strong initial syllables, suggesting that the strategy will be able to isolate most of the content words. The MSS has been explored in the context of various current models of word recognition (see, e.g., Cutler & Carter, 1987; Cutler & Butterfield, 1992; McQueen, Norris, & Cutler, 1994; Norris, McQueen, & Cutler, 1995) and is now seen by Cutler, Norris et al. as a separate component of a larger model, contributing along with other parts of the model, to segmentation behavior. Critically, Norris et al. (1995) have shown, using a word spotting technique, that competition between simultaneously active word candidates can modulate the size of the prosodic effects that prompted the original formulation of the MSS. The MSS has been incorporated into the SHORTLIST word recognition model (Norris, 1994): lexical hypotheses beginning with strong syllables are favored and their activation levels are increased, whereas the activation levels of candidates misaligned with strong syllables are penalized.

In the latest development of the research involving the MSS (Norris, McQueen, Cutler, & Butterfield, 1995), the MSS provides prelexical segmentation cues for the operation of a Possible-Word Constraint (PWC), which penalizes lexical hypotheses that leave a nonsyllabic residue between themselves and a clear syllable/word boundary. This constraint is implemented in the SHORTLIST model. We will not discuss the PWC further, below; we note only that the MSS remains intact within this formulation.

These developments in the modeling of the MSS have tended to eclipse the earlier concern with the distinction between open- and closed-class words (cf. Cutler & Carter, 1987; Cutler & Butterfield, 1992). In the SHORTLIST model, for instance, there is no explicit differentiation between the two word classes. The original concentration on the distinction between open- and closed-class lexica reflected the fact that a strong initial syllable is typical of open-class words, whereas closed-class words are typically subject to more phonological reduction (see, e.g., Cutler, 1993); indeed, in their calculations of the effectiveness of the MSS, Cutler and Carter explicitly assume that closed-class words are realized with weak initial syllables. However, using a representative sample of real conversational speech collected from 24 different speakers, Shillcock, Bard, and Spensley (1988) reported that some 32% of closed-class items were in fact pronounced with strong initial syllables.[1] This suggests that the prosodic generalization underlying the MSS may be of even

[1] In most cases the ''initial'' syllable is the only syllable for frequent closed-class words.

greater utility than is suggested in the earlier studies cited above, which were largely based on dictionaries or orthographic corpora. However, this increase in the potential of the MSS is somewhat offset by the finding in the same study by Shillcock et al. that around 11% of open-class words that are strong-initial in citation form were in fact pronounced with weak initial syllables.

We applied these results concerning strong syllables in real conversational speech to data from the CELEX database[2] and from a 460,000-word version of the London–Lund Corpus of English Conversation (LLC) (Svartvik & Quirk, 1980). The CELEX figures were calculated by weighting the word-type statistics according to spoken token frequency. These calculations showed that in conversational speech, if the MSS is defined simply as designating a strong syllable as the initial syllable of a word, then the MSS might be expected to identify some 50% of *all* word boundaries. This figure for all the words is composed of 31%, for the open-class words, and 19%, for the closed-class words, and reflects the fact that some 13% of open-class tokens were calculated as beginning with a weak syllable even in citation form.

These calculations reveal the MSS as a powerful bottom-up source of segmentation information for English. However, its utility must be measured against not just the boundaries it correctly predicts (''hits''), but also the occasions on which it makes an incorrect hypothesis about a boundary (''false alarms''). Our calculations show that some 5% of word tokens in speech will contain a noninitial strong syllable, as in *perceive.* This is therefore only a slight source of misleading information for the adult listener, even assuming that a morphologically revealing missegmentation like *per-ceive* is as uninformative for the processor as a morphologically unrevealing false alarm like *set-tee.* Even this source of missegmentation may be smaller for acquisition, given the relative infrequency of content words with weak initial syllables in the speech addressed to infants in the CHILDES corpus (MacWhinney & Snow, 1985). Finally, it should also be noted that the MSS needs to incorporate some additional assumption regarding syllable onsets so as to specify the exact input to lexical access where a syllable might legitimately start at the vowel or at more than one preceding consonant.

The MSS was originally developed as a model of adult behavior, but Jusczyk, Cutler, and Redanz (1993) have presented evidence that infants as young as 9 months, from English-speaking backgrounds, are sensitive to the contrast between words with a strong–weak metrical stress pattern such as *pliant* and those with a weak–strong pattern like *comply.* The test used was an infant head-turning paradigm, which measures the time for which attention

[2] CELEX Lexical Database of English (Version 2.5). Dutch Centre for Lexical Information, Nijmegen. The CELEX frequency counts are derived from spoken corpora totaling 1.4 million words.

is directed to a particular stimulus. Note that such experiments demonstrate only that infants are sensitive to metrical structure (or some other highly correlated property) and not that they actually use the MSS. In later sections, we will explore ways in which the MSS may be acquired. As Jusczyk et al. point out, the mechanisms for stress realization vary from language to language. In English, strong syllables tend to be higher in pitch than their weak neighbors (along with other distinctions such as longer duration and greater amplitude); however, in other languages these indicators are different. In Norwegian, for instance, stressed syllables tend to have a low pitch. It therefore seems that ''strong syllable'' is not necessarily a primitive based on inherent perceptual salience; rather, the category must either be learned or be selected as a parameter from a universal inventory of stress realization possibilities containing, at least, the syllable (French) (Mehler, Dommergues, Frauenfelder, & Segui, 1981), the strong syllable (English) (see, for example, Cutler, Mehler, Norris, & Segui, 1992), and the mora (Japanese) (Otake, Hatano, Cutler, & Mehler, 1993). However, from an empiricist viewpoint, to demonstrate that metrically based segmentation can be bootstrapped necessarily involves also bootstrapping the notion of strong syllable.[3] If we cannot take the strong syllable to be a primitive, then we need to look for other sources of segmentation information that could be used to reveal this prosodic generalization. We will provide below a strongly bottom-up account based on the low-level distributional regularities of English.

In summary, the MSS reflects a powerful generalization about lexical structure in English and is of central relevance to segmentation, possibly directly guaranteeing 50% of word boundaries using this bottom-up information. This assessment of its contribution is made in abstraction from the incorporation of the MSS in any particular model. We will demonstrate below how the relevance of the MSS may emerge in the course of computing simple distributional statistics at the subsegmental and segmental levels. First, however, we review the role of distributional approaches more generally.

Distributional Approaches with Orthographic Corpora

There has been considerable interest in distributional approaches to the segmentation problem, applied initially to orthographic, rather than phonological, representations of language. For example, Wolff (1977) analyzed corpora of written text with no word boundaries and built chunks from frequently occurring sequences of letters. Chunks could then be combined with letters or other chunks to create larger units. The construction of these chunks implicitly

[3] This is also true if we hypothesize that strong syllables are less frequent than weak syllables in English and that it is this markedness that makes them salient, along with their Norwegian counterparts.

segments the text, producing boundaries which are correlated with linguistically meaningful units. Redlich (1993) independently developed a related approach, justified on information-theoretic grounds. Brent (1993) studied segmentation of single orthographic words into morphemes, using a related technique based on the Minimum Description Length (MDL) principle.

These approaches are complementary to the method that we propose below. They operate by finding frequent sequences and assuming that these are linguistic units of interest. Parsing the input stream with reference to these units implicitly provides segmentation boundaries. Our approach is to discover segmentation boundaries directly, by searching for *infrequent* sequences. Our approach is, in a sense, lower level, since segmentation is prior to, rather than dependent on, the learning and storage of linguistic units.

More related to the approach we present below are the simulations by Elman (1990), in which a neural network is trained to predict the next element in a sequence which consists of a concatenation of ''words'' in random order. Error is high for prediction across word boundaries, and Elman notes that error can thus be used as a cue to segmentation. However, since distributional methods depend on the statistical structure of the natural language being learned, simple artificial examples can at best be suggestive.

Distributional Approaches with Speech Corpora

The use of distributional information to find word boundaries in speech has received much less attention in the psycholinguistic literature than the prosodic approach. Whereas Cutler and Norris have provided data showing that listeners use prosodic information in segmentation, less attention has been devoted to showing that other sources of information also contribute to segmentation.

The phonotactics of a language is the sequential constraints that operate on contiguous items at the segmental level. This sequential information can be described as a set of co-occurrence restrictions that hold within syllables. Thus, while the sequence /#pr/ is permissible in English, /#mp/ is not, where the symbol # represents a syllable boundary. Likewise, in syllable-final position, /mp#/ is valid, but not /pr#/. Some of these restrictions may have an articulatory basis, but this is not generally the case: /#mp/ for example would be a perfectly acceptable sequence in many Bantu languages. However, phonotactic constraints do not have to be absolute, they may simply be probabilistic: thus the sequence /nd/ is very common word-internally but much less common across word boundaries in English.

Empirical support for the infant's sensitivity to simple phonotactics comes from another infant head-turning experiment by Jusczyk, Friederici, Wessels, Svenkerud, and Jusczyk (1993) (see also Jusczyk, 1993b). In this study, there were two experimental conditions: in one, the infant was played recorded Dutch and English word lists; in the other the same material was played

after low-pass filtering (leaving only prosodic information). There was no difference in attention time given to Dutch and English when only prosody was present. However, in the unfiltered speech condition 9-month-old infants attended significantly more to their native language (but 6-month-olds did not). So it seems that the high-frequency information (the identities of segments as opposed to suprasegmental information) was responsible for the effect. This result is supported by Friederici and Wessels (1993) with intra- as opposed to interlinguistic stimuli. Therefore, it seems that during the first year of life the infant becomes sensitive to phonotactic statistics; also over this period some nonnative contrasts in phonological space are being merged (see Werker, 1993, for a review).

Brent and colleagues (Brent & Cartwright, 1996; Cartwright & Brent, 1994) demonstrate the utility of supplying a model of segmentation with categorical phonotactic information consisting of all the legal initial and final consonant clusters found in the English lexicon. Such a filter on the acceptability of segmentations adds substantially to the effectiveness of a data-driven approach based on the MDL principle, as did the addition of a principle requiring each word and syllable to contain a vowel, when tested on sections of the CHILDES corpus that had been given a phonological transcription. In contrast, syllable-internal constraints reflecting the sonority hierarchy were not found to be effective. This research is attractive because it demonstrates the potential of assessing the independent contributions of formally distinct aspects of phonotactic information, but the fact that the phonotactic information was simply supplied to the model limits the implications of this research for theories of the development of segmentation. Also, the use of categorical phonotactic information (legal *versus* illegal), as opposed to richer, probabilistic information about word boundaries, may have led to an underestimation of the value of phonotactic information.

Weakly Bottom-up Phonotactic Approaches

Working within a speech engineering paradigm, Harrington et al. (1988) developed a trigram model in which word boundaries were included in the symbol set as a method for deciding on the plausibility of boundary insertion. Using a 23,000-word lexicon they extracted the set of all possible phoneme triples permissible either word-internally or across word boundaries. Their segmentation algorithm then used the information about which sequences were impossible word-internally to place boundary points in a deterministic fashion. Using this primitive method they reported finding 37% of the boundaries in a test-set of sentences with a hits:false-alarms ratio of about 8:1 (a hit is the detection of a real boundary, a false-alarm is the positing of a spurious boundary). However, the original 23,000 word lexicon was reduced by eliminating compounds and inflected forms, leaving only 12,000 items. This was done to stop undersegmentation which would arise from licensing

forms such as /mʌnθs/ with the word-internal trigram /nθs/ which would prohibit desirable segmentations such as *month seems.* This simplification, coupled with (a) the use of a dictionary with no frequency information and (b) testing with neatly formed sentences atypical of genuine conversational English, makes the efficacy of this approach hard to determine for real data. The authors were unconcerned with developmental issues and the model is only weakly bottom-up, since the trigrams encode information about known word boundaries, and thus is only useful in describing adult behavior.

A similar model that seeks to exploit prior knowledge about the distribution of word boundaries in a phoneme stream to postulate segmentation points has been suggested by Vroomen and van den Bosch (1992). They trained a supervised neural network to respond when a word boundary was present in a phonemic input stream with no word boundary marking. Their model is quite successful as a description of adult behavior, but once again their methodology is incapable of addressing the bootstrapping question since the model is only weakly bottom-up: training involves telling the network when a segmentation point is present.

Strongly Bottom-up Phonotactic Approaches

One seemingly strongly bottom-up model that used phonotactics, an interesting historical precedent for our approach, was proposed by Harris (1955) in which a discovery procedure for morphemic structure is described. The central idea of this work was to deduce the location of boundaries in a continuous phoneme stream by calculating the number of phoneme types that were possible candidates for continuation at each point in a phrase (see Fig. 1). This number is referred to as the *successor count* of a string. Segmentation points are then postulated just after a peak in the successor count.

The intuitive explanation of the viability of this technique is that within words and syllables, the sequencing of phonemes is more constrained than at word boundaries. For example, after the phoneme sequence /kar/, if there is no word boundary, then there are only a handful of possible successors including /p/, /t/, and /z/. However, if there is a word boundary, then practically any phoneme can continue the sequence.

Using this technique Harris stated that he was able to obtain convincing morphemic segmentations of phrases, although he did not provide any quantitative results. However, the technique is not without its drawbacks; one of the most obvious is the likelihood of oversegmentation in the case of lexemes which are themselves initial portions of other words. Thus *carpet* will induce segmentation after *car-* since at this point the possible word ending licenses a large number of possible successor phonemes. Note that exactly the same problem plagues any strictly left-to-right algorithm which has no ability to ''look back.'' Harris manages to control for this problem by augmenting the

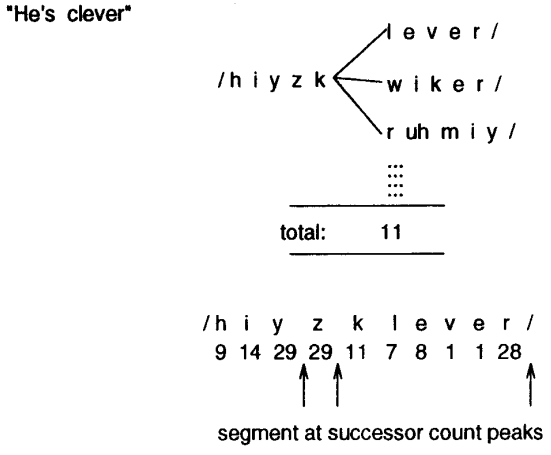

FIG. 1. Segmentation method used by Harris (1955). At each point in a given string of phonemes, the number of different phonemes that can be used in continuing the phrase was established. In this example possible continuations include /l/ (*he's clever*) and /w/ (*he's quicker*). Boundaries are placed after peaks in this successor count.

algorithm with predecessor counts and insertion counts, obtaining even better results.

Harris was not explicitly concerned with developmental issues, the discovery procedure being an abstract linguistic tool. However, if one were to apply the model to development, its success would be hindered by the fact that successor count information was not extracted from real data. The successor counts of a string were established by elicitation of possible completions from introspection by informants. Therefore, it is hard to argue that the data used provide a fair representation of the problem as faced by a human learner. Furthermore, the algorithm was applied only to coherently formed sentences, untypical of real speech, with the informants having access to all of the earlier part of the phrase being presented. For these reasons Harris' results do not constitute a demonstration that the bootstrapping of segmentation is possible using phonemic information alone.

n-GRAM MODELS

Our major criticism of the above bottom-up techniques for uncovering segmentation points is that they have not demonstrated their efficacy with real data that are representative of the input that confronts human listeners and learners. Because of this problem we now present what we believe to be much more adequate input for any model of segmentation: a large corpus of real speech. A detailed account of the corpus is not given here, but see the description by Shillcock, Hicks, Cairns, Levy, and Chater (in press). Use of

this data will allow us to make much more substantive claims than is possible using small-scale, unrepresentative input.

A Phonological Retranscription of the London–Lund Corpus

The London–Lund Corpus is a large body of English conversation transcribed orthographically and available on-line. Because of its size, an automatic method was developed for its phonetic transcription. First, the words were replaced by their phonemic citation forms using an on-line dictionary. Then these forms were input to a set of rewrite rules that introduced phonological alternations into the string, such as assimilation and vowel reduction. None of the rules used word boundary information to specify its context of application. The output from the rule set was a corpus of some 1.5 million phonetic segments.

It is, of course, impossible to recreate the original speech data, but this method has two main advantages. First, we need a very large corpus of conversational speech if its statistics are to be representative; at present there is no sufficiently large corpus with a genuine phonological transcription. Second, this method provides a higher-order approximation to genuine data when compared, for instance, with a corpus derived from a phonemic dictionary in combination with word frequency counts. Thus, our data are representative of the distribution of strings of closed-class words such as *if I can.* As already emphasized, any adequate model of segmentation must cope with such input.

There are two important characteristics of our corpus. First, there is no explicit marking of word boundaries. All rules for coarticulation apply equally inter- and intralexically. Second, 17.0% of words in the LLC occur after a pause or speaker changeover, so pausing/changeover is a good cue to segmentation. All pause/changeover markers were removed to avoid instantiating pause as a phonemic category. Because of these facts, the data represent a ''worst case'' for testing models of segmentation, in that if segmentation is possible with these data, then the inclusion of pauses and some phonetic/acoustic cues can only serve to improve performance.

Weakly Bottom-up n-Gram Models

Although the main focus of this paper is on models that are compatible with a theory of acquisition, we will describe the results of a nondeterministic extension to the model described by Harrington et al. (1988), since these results clarify the parameters of the segmentation problem.

Bigram models. Using the whole corpus described above, but before word boundary markers were removed, the prior for all bigrams: $\{\langle p^1, p^2\rangle \mid p^1, p^2 \in P\}$ was calculated, where the pair either was word internal or straddled a word boundary (here we use set theoretic notation, read as ''the set of all phoneme sequences $\langle p^1, p^2\rangle$ such that both phonemes are members of the set

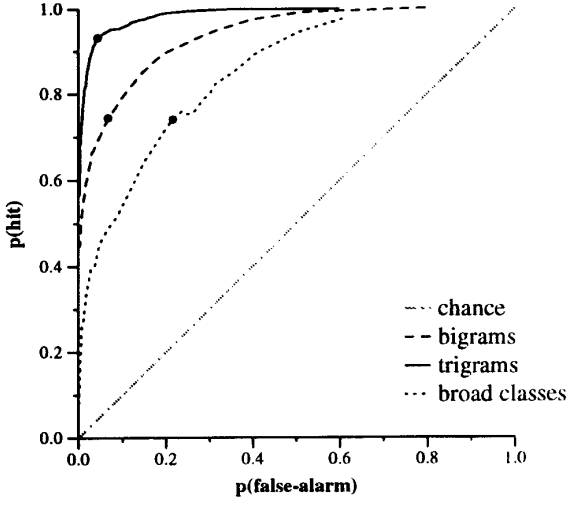

FIG. 2. ROC (Receiver Operating Characteristic) graph for *n*-gram segmentation performance, where the corpus which contributed the statistics was marked for word boundaries. The different curves show the results for bigrams, trigrams (with one-phoneme word constraint, see text), and a trigram model where the transcription was in terms of the six broad phonetic classes. The points of maximum mutual information are marked with a dot on each curve.

of all phonemes *P*.'' Whereas Harrington et al. simply inserted boundaries in sequences of phonemes which were impossible word-internally, we can use the ratio of the prior that a pair $\langle p^1, p^2 \rangle$ occurs across a boundary to the prior that it occurs within a word, denoted $p_{across}(\langle p^1, p^2 \rangle)/p_{within}(\langle p^1, p^2 \rangle)$ to decide when to propose a boundary. When this ratio rises above a certain cutoff point we insert a boundary. When the cutoff is set high, the performance of this model tends toward the basic Harrington et al. deterministic model.

The results of running this segmentation algorithm on a 10,000-phoneme (approximately 2700-word) test stretch of the same corpus can be seen in Fig. 2, where we plot the probability of a hit versus a false-alarm as the cutoff is varied. Selecting different cutoff points can provide a performance such as the detection of 45% of the boundaries in the test stretch with a hits:false-alarms ratio of 45:1, or 66% of all boundaries with a hits:false-alarms ratio of 9:1 (recall that Harrington et al.'s model detected 37% of the boundaries at a hits:false-alarms ratio of 8:1, although note that the respective test corpora are not comparable). Note that in Fig. 2 the hits and false-alarms are plotted as probabilities, whereas the hits:false-alarm ratios are calculated from absolute counts (and the hits:false-alarm ratios may not therefore be read directly from the ROC curve). The probabilities are calculated using the fact that there is a mean of 3.7 phonemes per word in the LLC.

Where exactly to place the cutoff point is a question that depends on our theory of how much of a problem false-alarms and misses pose for the human speech processor, which will reflect assumptions about processor modularity, parallelism in processing, and so forth. The cutoff point may also be susceptible to changes in the nature of the input and the availability of other types of information relevant to segmentation. At this stage in the investigation of segmentation behavior, the ROC curve is the best representation of the potential contribution of different types of distributional information to segmentation. However, one can measure how well the segmentation algorithm performs in a pretheoretical manner by taking an information theoretic measure such as *mutual information* at each cutoff point and choosing the cutoff at which this measure is maximized.[4] In effect, the mutual information measure tests whether the general shape of the distributions of boundary points is the same for the segmentation algorithm and the veridically segmented corpus and also the extent to which the individual decisions match. At the mutual information maximum of 0.24 the detection rate is 75% with a hits:false-alarms ratio of 4.7:1. The maximum mutual information points are marked on the graph.

In conclusion, simple distributional statistics for segment bigrams seem to offer information of substantial relevance to speech segmentation. We now look at trigram models to explore more of the potential of *n*-gram models, before discussing the reliance of such models on a finegrain phonemic transcription.

Trigram models. Further improvement can be made on the performance reported above by using trigrams rather than bigrams. We collected the priors of all triples: $\{\langle p^1, p^2, p^3\rangle | p^1, p^2, p^3 \in P\}$ that were word-internal, had a boundary between p^1 and p^2, or had a boundary between p^2 and p^3. However, now we have two ratios p_{across}/p_{within}, where p_{across} can correspond to the sequence $\langle p^1, \#, p^2, p^3\rangle$ (henceforth denoted P#PP) or $\langle p^1, p^2, \# p^3\rangle$ (PP#P). As a first step, we simply took the mean of the two ratios and moved the cutoff point relative to this figure. A further complication that arises through the use of trigrams is the tendency to oversegment when there are one- and two-letter words in the input. Consider the boundary between the first two

[4] The mutual information of two sources, $M_{S,T}$ is defined as follows: $M_{S,T} = I_S + I_T - I_{S,T}$, where I_S and I_T are the total information of sources S with states s_i and T with states t_i, respectively, and $I_{S,T}$ is the joint information between S and T. Thus

$$I_S = -\sum_i p(si)\ log(p(si))$$

$$I_T = \sum_i p(ti) log(p(ti))$$

$$I_{S,T} = \sum_{i,j} p(si,tj) log(p(si,tj)).$$

For binary data such as ours, each source has only two states (corresponding to *boundary-present* and *boundary-absent*) yielding four possible combinations which correspond to *hit, false alarm, miss,* and *correct rejection.*

words in /taɪpɪŋɪtʌp/ (*typing it up*). The trigram ⟨ŋ, ɪ, t⟩ does not occur word-internally, but does occur as P#PP; therefore the p_{across}/p_{within} ratio here is infinite, and hence we insert an obligatory boundary after /ŋ/. However, the sequence ⟨ɪ, t, ʌ⟩ also never occurs word-internally, so we also insert a boundary after /ɪ/ yielding the segmentation /taɪpɪŋ#ɪ#tʌp/. A remedy for this problem is to have a list of permissible one-phoneme words (for present purposes just /ə/ and /ou/) and not to license segmentations that create one-phoneme words not on this list; this constraint resembles the Possible Word Constraint suggested by Norris et al. (1995). Having done this, the results for the segmentation of the same test stretch of corpus as before are shown in Fig. 2. The trigram results show a considerable improvement on the bigram figures, with performance ranging from detection of 57% of the boundaries with a false-alarm rate of 65:1, to the mutual information peak of 0.418 with 93% detection at a hits:false-alarms ratio of 9:1.

The algorithm does indeed show some oversegmentation of inflectional forms as Harrington et al. (1988) realized would happen (recall that they removed all inflected forms from their dictionary, considering the resulting oversegmentation of forms such as *three # month # s # time* to be a situation from which recovery was possible using morphological rules). However, as can be seen from the current results, these cases are really quite rare in normal conversational speech. Another error is to oversegment words which begin with a weak vowel, as in /tɔk#ə#baut/ for *talk about*, though once again such cases are rare.

This success of bigram and trigram models in segmenting the transcribed speech stream relies on a detailed and unambiguous phonemic input string, something which may not be obtained in real human listening. In real speech, phonemes are realized with numerous variations in both time and quality. Phonological reductions such as /fnɔlədʒi/ for *phonology*, where the initial /fn/ is not permitted in more articulated speech, would cause undersegmentation if included in the training set or oversegmentation if not included. However, the size of these problems can be ascertained only with reference to their frequency of occurrence in real data.

We carried out one test of the reliance of these results on fine-grain transcription with a full phonemic inventory, by using the six broad phonetic classes employed by Zue and colleagues (see, e.g., Huttenlocher and Zue, 1983).[5] We retranscribed our corpus in terms of these six classes and then constructed a trigram model using the same procedure as before. Not surprisingly, we found the performance was considerably reduced when compared to the fully transcribed trigram model (see Fig. 2). However, although weakened, performance is still good, with a mutual information peak of 0.123 with 74% detection at 1.5:1 hits:false-alarms ratio.

[5] The classes are defined as follows. Stop: p t k b d g tʃ dʒ. Nasal: m n ŋ. Weak fricative: f θ v ð h. Strong fricative: s ʃ z ʒ. Liquid or glide: l r j w. Vowel: all vowels.

In conclusion then, these results show that *n*-gram models can make a substantial contribution to solving the segmentation problem, although their performance is weakened somewhat by a poor phonetic representation. In contrast, the MSS requires a phonetic distinction to be made only between strong and weak vowels and hence will be relatively unaffected by an impoverished phonetic representation, if the decision as to precisely where a syllable starts can still be made accurately. In a fuller model of speech segmentation, *n*-gram statistics, possibly of more variety than the representatives described here, might appear as one source of constraint used in conjunction with other information, such as the MSS.

Strongly Bottom-up n-Gram Models

We shall now consider some purely bottom-up, unsupervised, *n*-gram models: systems in which no word boundary information is present during training, which can therefore be considered as better candidates for models of development than the weakly bottom-up models considered above. The results for these strongly bottom-up models should be poorer, compared with those given above for the weakly bottom-up models, reflecting the additional constraint that no veridical word boundary information is available. All the models described below were tested with the same corpus as input as those above, but with no explicit word boundary information at any stage.

Successor counts and perplexity. The first model we describe involved segmenting at a successor count peak in exactly the same fashion as Harris. Note that the successor count is defined over phoneme types, not tokens. Therefore, this naive technique does not take frequency of possible successors or successor bigrams into account. A ''peak'' was defined by normalizing the successor counts for phoneme type and cutting off at an arbitrary point above the mean. The results of this minimal method were very poor. It did not perform significantly above chance. In other words, no matter where the cutoff point was placed above the mean, there were always some three false-alarms for every hit when segmenting a 2700 word test stretch of the corpus. Since the average length of a word in this corpus is 3.7 phonemes, if a boundary was simply placed after every fourth phoneme, then the result would be a hits:false-alarms ratio of about 1:4. Thus, this instantiation of the Harris technique confirms that it is not viable for use with a corpus.

A more sophisticated model can be constructed by taking into account the frequency of the trigrams, thus defining the successor count over tokens rather than types. We used the information theoretic definition of ''perplexity'' in a symbol string to provide a continuous value that could be thresholded to yield lexical boundary postulates.[6] Although taking this step improves on the

[6] The perplexity of a two-phoneme string $\langle p_1, p_2 \rangle$ is defined as:

$$\mathrm{Per}(p_1, p_2) = \sum_i p(\langle p_1, p_2, p_i \rangle) \log p(\langle p_1, p_2, p_i \rangle) \mid \mathrm{p}_i \in P.$$

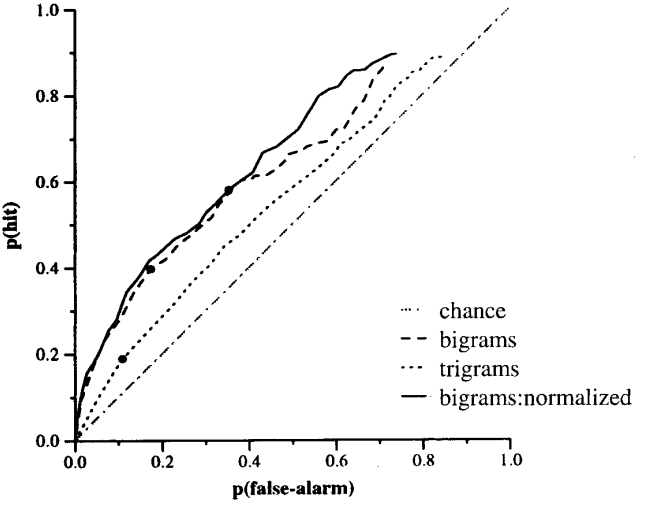

FIG. 3. ROC graph for *n*-gram segmentation performance (corpus with no boundary marking).

simple type-based successor count, it still does not produce results that merit further exploration of this paradigm.

We therefore conclude that perplexity measures are unlikely to be of significant use in segmenting speech, at least when they are defined over triples, although it is still possible that higher order *n*-tuples would improve performance. In addition, the picture may be different in the supervised case in which the location of word boundaries is known during the model's training. We now turn to another class of strongly bottom-up model that is extremely simple and more effective than approaches using successor counts and perplexity.

Simple n-gram models. The intuition behind the class of models we now describe is that sequences which are highly frequent will tend to identify word-internal chunks of the phonetic string and that segmentation points should therefore be proposed at points where the probability of a particular sequence is low. Again we collected the probabilities of all strings $\{\langle p^1, p^2\rangle | p^1, p^2 \in P\}$ but this time on the version of the corpus that contained *no* word boundary marking. We segmented the 2700 word test corpus by placing a boundary in the string at points where $p(\langle p^1, p^2\rangle)$ was below a certain threshold. The results are shown in Fig. 3.

The mutual information peak of 0.026 is at 38% boundaries detected with a 0.85:1 hits:false-alarms ratio. However, a lower cutoff produces poorer, but more reliable performance, with 24% detection at 1.25:1. We found that constructing an identical model, but using trigrams, produces results that are in fact much poorer than those for bigrams (see Fig. 3). This is because the

trigrams cannot pick out the one- and two-phoneme words that form a substantial part of the input (in fact 37% of all word tokens in our corpus have either one or two phonemes). However, by normalizing for phoneme frequency (i.e., using the probabilities p ($\langle p^1, p^2, p^3 \rangle$) that are conditional on some or all of the constituent phonemes) some improvement is obtained. This improvement occurs because one- and two-phoneme words are principally function words, and hence the most frequent words with the most frequent phonemes, so weighting against these cases yields better performance. However, the improved trigrams do not out-perform the bigram results, less so when the bigrams are normalized for the identity of the two phonemes (see Fig. 3). Intuitively, the reason these *n*-gram models are successful while their perplexity-based counterparts are not is that here we are in effect matching expectations against reality, while the perplexity model is purely predictive—the estimated perplexity of a string $\langle p, t \rangle$ is never matched against the actual probability of its completion $\langle p, t, ə \rangle$.

It appears, then, that sensitivity to phonotactic statistics would provide a neonate with some ability to break up the input speech stream. Clearly, this information alone will not provide a complete segmentation of speech input, but its use will allow an initial purchase on segmenting the continuous multiword speech stream.

Problems with n-Gram Models

All the *n*-gram models so far discussed share the principle that sequential constraints are expressed at the phonemic level. Two criticisms can be made of this assumed input. First, since we intend our model to predict the very first stages in the process of extraction of lexical chunks by an infant, if we use categorial *n*-grams, then this necessarily postdates the beginning of categorial phonemic perception in an infant (cf. Kuhl, 1983). Second, despite the descriptive convenience afforded by the category ''phoneme'' and the role of a phonemic level in models of spoken word recognition such as TRACE (McClelland & Elman, 1986), accounts have been advanced more recently, both for formal phonological description (see, e.g., Harris & Lindsey, 1993) and for psycholinguistic models of processing (see, e.g., Norris, 1990; Gaskell & Marslen-Wilson, 1995), where phonological features are mapped directly into words, and any perception of segments may be orthogonal to word recognition. An account of speech segmentation based on an input composed of phonological features may be both more parsimonious and more psychologically realistic. Indeed it may be the case that phonotactic predictors of segmentation are present in combinations of subsegmental features: an example is the ''sonority hierarchy,'' a linguistic principle which states that the more peripheral a segment in a syllable, the less likely it is to be sonorous (''voicey''). Thus, syllable nuclei contain highly sonorous vowels, while voiceless fricatives normally occur only in syllabic extremities (for instance,

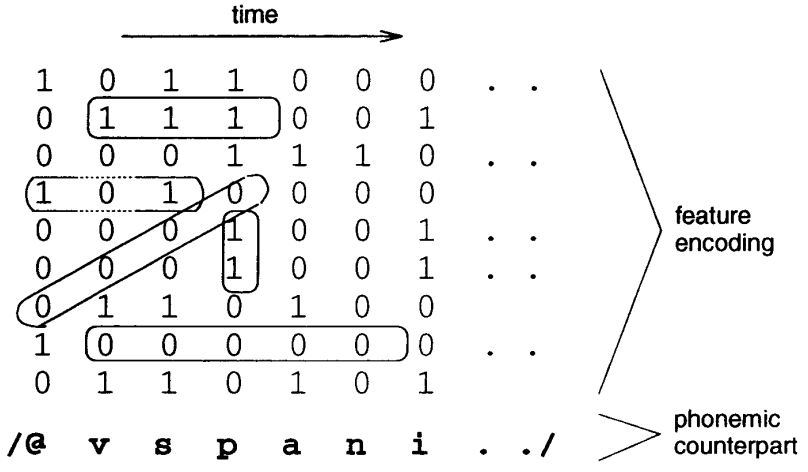

FIG. 4. A sufficient encoding device for feature input must be able to capture both syntagmatic (horizontal) and paradigmatic (vertical) dependencies.

freaks shows a low–high–low sonority profile). Therefore it would be advantageous to segment when sonority is low, a prediction tested by Brent et al. (1996) who in fact found only a nonsignificant effect of this particular constraint. We explore, below, one model from the class of models that can acquire phonotactic information, but that use feature descriptions. As feature input arrives over time, such models must be sensitive to both syntagmatic and paradigmatic relationships between features, where the dependency is continuous or possibly discontinuous (see Fig. 4). Use of a feature-based model will allow us to address the variability of speech in terms of allophonic variation, but still ignores time variation in phoneme realization. In the *n*-gram models discussed above, time is divided into discrete timeslices, with one segment per timeslice. Feature-based models allow individual features to be smeared across time, approximating the coarticulation and coloring of sounds found in real speech (see Gupta and Mozer, 1993, for such an input representation). In the model we describe below we will incorporate coarticulation rules, but we abstract away from any spreading of features in the input.

It would also be advantageous to have a system that is flexibly self-selective in terms of the lengths of input sequences that affect its output responses. Using a model that effectively mixed *n*-grams for various *n* would potentially allow us to sidestep the oversegmentation problems that we have encountered when using trigrams. With these motivations, we now describe a connectionist network for modeling segmentation that uses feature input and is not fixed to any particular cardinality of *n*-gram. As input to this feature-based model, we translated the corpus described above into a nine-bit binary feature vector

representation where the features are taken from the Government Phonology scheme of cognitive elements (see Harris & Lindsey, 1993; Kaye, Lowenstamm, & Vergnaud, 1985; Shillcock, Lindsey, Levy, & Chater, 1992). Williams and Brockhaus (1992) have shown how the Government Phonology elements can be automatically extracted from the speech stream, so we have reason to believe that coding in this way represents a step further toward ecological validity.

A CONNECTIONIST NETWORK FOR MODELING SEGMENTATION

To conduct segmentation using a feature-level description requires that we have some way of calculating the probability of the occurrence of each feature at a particular point in the sequence, given the feature vectors of the preceding segments. A successful predictive method should uncover the sequential statistics which are defined at the feature level. Only to the extent that our method for prediction picks up within-word phonotactic regularities will the failure of such regularities to hold across word boundaries be informative for segmentation.

In abstract statistical terms, this kind of prediction problem is naturally modeled as a problem of regression. Standard linear regression is, however, unlikely to be adequate, since phonotactic constraints between features are highly variable depending on context. A more promising approach is to use the powerful and general nonlinear regression methods implicit in back-propagation learning using a neural network (see White, 1992).

Network Architecture

A flexible architecture for tackling prediction problems is the simple recurrent network (SRN) (Elman, 1990; Norris, 1990), which comprises a standard feed-forward neural network, augmented with a set of ''copy-back'' or ''state'' units that permits a limited feed-back within the system. In operation, the activation values of a layer (typically the hidden layer) are copied-back onto the context units, on a one-to-one basis. If the context units are then connected to the hidden layer with standard feed-forward modifiable connections the network will have access to a ''memory'' of previous hidden unit states and can respond to constraints which may in principle be defined over any number and combination of previous inputs, over any time period. In practice, learning is generally much more successful when constraints are relatively local. SRNs have been used productively in modeling a range of aspects of language processing (e.g., Cleeremans, Servan-Schreiber, & McClelland, 1989; Elman, 1990, 1991; Norris, 1990, 1993; Shillcock, Levy, & Chater, 1991; St John & McClelland, 1990).

SRNs can be trained using standard back-propagation, since their feedback connections are not modifiable. The back-propagation learning algorithm was developed for feed-forward networks, where it can be shown to perform

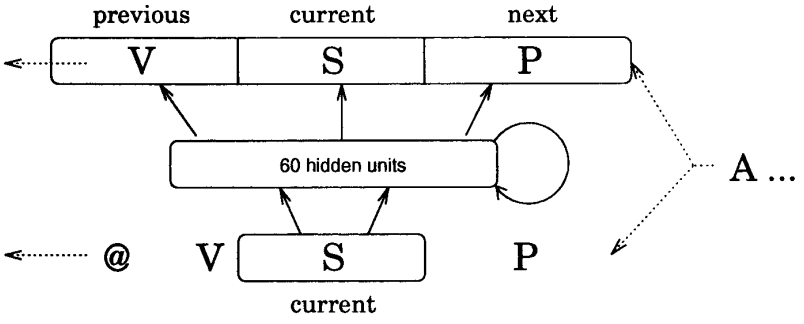

FIG. 5. The network architecture. The solid arrows between layers indicate complete connectivity with modifiable unidirectional links. The dotted arrows show how the input corpus arrives over time to specify the input and output target. (The symbols are from a machine-readable phonetic alphabet.)

gradient descent in error space—that is, the weights are changed slightly in the direction which reduces error as much as possible. For SRNs, the copy-back method is computationally cheap, but computes only a rough approximation to true gradient descent. Since we are interested in obtaining the best prediction that we can, we used the computationally more expensive method, ''back-propagation through time'' (BPTT; see Rumelhart, Hinton, & Williams, 1986) which allows the error signal to be back-propagated through longer stretches of time than in the SRN (see Chater & Conkey (1992) for a detailed comparison).

Network Training

The network has a recurrent, self-supervised architecture (see Fig. 5). We use ''self-supervised'' to mean that the input and target output for network training are specified from a single data stream, given a three-place buffer. The task is to echo the current slice of input, to remember the previous, and to predict the next. Providing additional tasks has been found to improve performance in training SRNs (cf. Abu-Bakar & Chater, 1993; Maskara & Noetzel, 1992). Input is from the Government Phonology transcription of the corpus described above. Noise is added to the input by flipping features from 0 to 1 (or vice versa) with a certain probability, in order to encourage the network to rely on sequential information (i.e., if the identity of the current segment is ambiguous or unclear, then the net will have an incentive to use the local phonetic context to recover its identity). The net is trained using BPTT, a steepest descent procedure, and a cross-entropy error measure (see Hinton, 1989); cross entropy is a good measure to use to interpret continuous valued outputs as probabilities of binary decisions). Training comprises two passes through a training stretch of the corpus 1 million segments in length

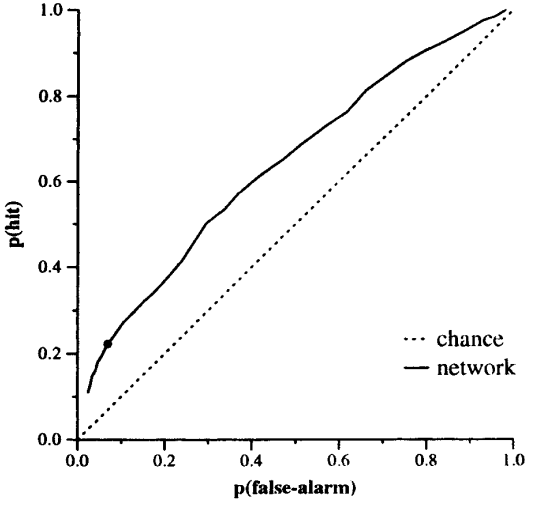

FIG. 6. ROC graph for network segmentation performance.

(with different noise on each pass), thus 2 million segments in total. The learning rate is decayed as training progresses.

Network Segmentation

The rationale used in postulating boundary points is that high perplexity makes the next segment difficult to predict. If prediction is hard, then error will be high. Thus, boundaries are proposed at peaks in the error score on the prediction output units (marked ''next'' in Fig. 5).

The model was tested by providing as input a noise-free 10,000-segment (approximately 2700 words) stretch of corpus and measuring the cross-entropy error on the prediction subgroup of the output units. This yields a variable error signal in which we define a ''peak'' by placing a cutoff point at varying numbers of standard deviations above the mean. The effects of choosing increasingly more stringent cutoff points can be seen in Fig. 6, where we plot how the hit and false-alarm rates vary with the cutoff point.

The results for segmentation using the cutoff that maximizes the mutual information at 0.023 are shown in Fig. 7. At this cutoff point 21% of the boundaries are correctly identified with a hits:false-alarms ratio of 1.5:1. Because calculation of the *a priori* probability of *random* segmentation performance is complex, we evaluate the significance of these results by comparison with a random segmentation algorithm which was averaged over five different runs. This algorithm was designed to yield a distribution of chunk length similar to that of the network. In other words, if the network produces pseudowords that are of length two 30% of the time, length three 20% of the

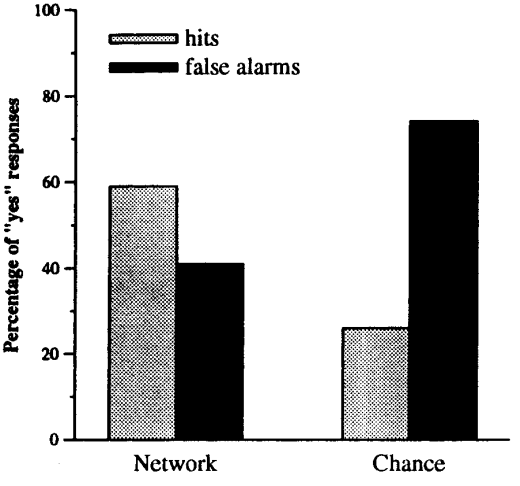

FIG. 7. Segmentation behavior of the network compared with chance behavior.

time, and so on, then we tailor the random algorithm so that it produces the same distribution of pseudoword length. This is a more stringent test of the network's performance than comparison with a random segmentation algorithm that uses a uniform distribution.

Although network performance peaks with correct identification of about one in five boundaries in the test corpus, there is a sizable proportion of false alarms at this cutoff (i.e., cases in which the network predicts a boundary when in fact there is none). It may well be that although the false alarms do not actually correspond to existing boundaries in the test stretch, they are actually plausible guesses based on the low-level data that are the only information source available to the model. We tested this hypothesis by examining the phonological acceptability of the boundaries that the model postulates.

The syllable-initial and syllable-final trigrams from a phonemic dictionary provide a simple measure of phonotactic well-formedness. If a boundary postulate creates an initial or final trigram that is present in the dictionary then it is categorized as an acceptable guess; if not it is said to be malformed. Thus, if the network posits a boundary before /sp/, we check to see if the trigram ⟨#, s, p⟩ is present in the dictionary. Figure 8 shows the proportions of phonotactically malformed boundaries for both the network and the random segmentor in syllable-final and syllable-initial positions. Only those boundaries that were false-alarms were included (by definition the hits are well-formed). The network's performance is far superior to that of the random segmentor: for the initial boundaries, $\chi^2_{(1)} = 221.8$, $p < .001$; for the final boundaries, $\chi^2_{(1)} = 119.1$, $p < .001$. Of all the network's boundary postulates (hits and false-alarms) 80.8% are phonotactically well formed.

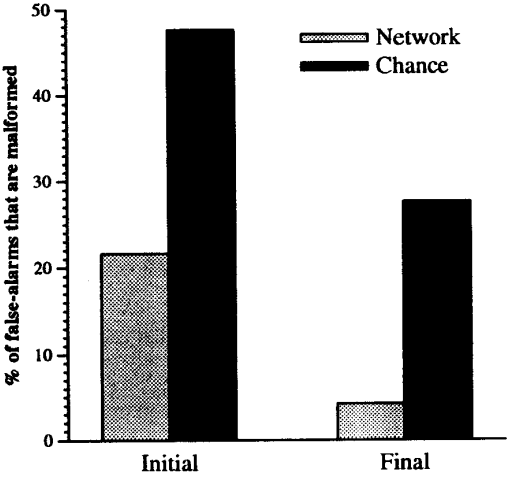

FIG. 8. Segmentation behavior of the network, showing the percentage of incorrect responses that are malformed in syllable-initial and syllable-final positions.

Our use of ''phonotactically well formed'' in the previous paragraph is nearly synonymous with ''syllabic'': boundaries which are phonotactically well formed are necessarily possible syllable boundaries. Given that in the LLC around 77% of syllables are word-initial, a device that correctly inserted syllable boundaries would also locate word boundaries 77% of the time. Given this, since 80.8% of the network's boundary postulates are syllabic, one would expect that 62.2% of boundary postulates would correspond to word boundaries. In fact, 59.3% of all the boundary postulates inserted by the network are actual word boundaries (this difference is not significant: $\chi^2_{(1)} = 1.613$, n.s.).[7]

These results show that the network's modest performance in finding lexical boundaries does not reflect a distinction between word and syllable boundaries; detecting syllable boundaries is sufficient, given the relationship between syllables and words in spoken English. Our results indicate, surprisingly, that phonotactic constraints are not stronger over syllable boundaries as opposed to word boundaries in spoken English. If constraints over syllable boundaries were stronger than those over word boundaries, then prediction would be easier, and network error would be lower, licensing fewer segmentations. Note that the set of lexical boundaries is not simply a subset of the set

[7] A precursor of this aspect of the investigation is found in work by Elman and Zipser (1988), who used a network and an identity mapping procedure to analyze relatively raw speech data and reported that ''hidden unit patterns were apparently associated with syllable onsets'' (p. 1625).

of syllabic boundaries: in connected speech, resyllabification can occur across word boundaries, thus *I bit him* may be represented as /ə#bɪ#tɪm/.

In summary, the network's strongly bottom-up performance is modest, at around one-fifth of word boundaries, and is coincident with syllable boundaries, but represents an initial purchase on the segmentation problem. We now consider how pause information may augment this performance.

Using Pause Information

The performance of this basic model is made more realistic by adding in the pauses and speaker changeovers that we removed when the corpus was retranscribed. We assume that a pause is an unambiguous cue to a lexical boundary and that therefore this information could be used by our strongly bottom-up system without fault. In our test corpus there are approximately 2700 words and 470 pauses (17.3%). Of the 521 initial boundaries that the model identifies (when the cutoff is the mutual information maximum) 134 of these occur after a pause (25.7%). This is significantly higher than chance ($\chi^2_{(1)} = 13.07$, $p < .01$), so the network's decisions are somehow correlated with pause location. The most likely explanation for this is the fact that the net tends to segment more before open-class words, a behavior to which we return below. If we add the number of pauses that the network does not detect back into the hits total for the network, then performance is considerably improved: now the system will detect 32% of the boundaries in the test stretch of the corpus with a hits:false-alarms ratio of 2.4:1.

Evaluation of Network Performance

The network's performance in segmenting the test corpus revealed a substantial false-alarm rate (at least when the model is operating optimally, though this level may be reduced if a lower detection rate is selected) and miss rate, although the real extent of the problem can be ascertained only with reference to a complete model of lexical access. Three points can be made about this level of performance. First, as shown above, although many false-alarms do not correspond to existing boundaries in the test stretch, they are actually plausible guesses based on the low-level data that are the only information source available to the model, preserving English syllabic structure. Second, with reference to false-alarms, some oversegmentation is arguably better than undergeneration in adult lexical access (but not in acquisition); indeed, Cutler and Butterfield (1992) show that adults tend to oversegment before strong vowels in slips of the ear. If the processor suggests boundaries on the basis of acoustic information, then higher processing can still cancel the boundary hypothesis without compromising processor modularity. However, if no boundary is proposed bottom-up, system modularity must be infringed upon to correct the mistake: in a purely bottom-up model, it is

possible to stop lexical access if initiated spuriously, but not to initiate it if a boundary point has been missed by bottom-up detectors.[8] Third, the child's lexicon may not contain adult-like entries (Mehler, Dupoux, & Segui, 1990; Menn & Matthei, 1992). In terms of the emergent lexicon, segmentation misses will be relatively inconsequential; meaning can often still be attributed to combinations of words, as in /əmin/ (*I mean*) which our model tends not to segment, but this is not typically so for parts of words. Storage will be attempted for multiword strings (MacWhinney, 1978; Peters, 1983), which will be relatively unsuccessful for longer strings. In cases in which storage was initially successful, the representation will simply not be reinforced by subsequent exposure.

Network Performance and the MSS

If phonotactic information and prosodic information are both candidates for cueing segmentation, then we may ask if the two sources of information tend to complement one another by isolating boundaries that are qualitatively dissimilar. Below, we will compare the output of our model with that of the MSS. Further, we will suggest that there is a statistical basis for the emergence of the MSS in our purely bottom-up model.

We investigated the performance of the model by counting the instances in which a boundary is correctly postulated before a strong or weak syllable. The definition of ''weak'' and ''strong'' is not trivial however. Although the presence of schwa (/ə/) invariably produces a metrically weak syllable, other short vowels such as /a/ and /ɪ/ can be either full or reduced depending on context (the version of the corpus used was not transcribed with metrical markings). As an operational definition of ''weak'' and ''strong'' syllables we took the lax vowels /ə/, /ɪ/, and /ʌ/ to indicate weak syllables, and all other monopthongs and diphthongs to indicate strong syllables. Because some of the instances of /ɪ/ and /ʌ/, which we classify as producing weak syllables, will actually produce strong ones, if anything these definitions will tend to artificially boost the number of weak syllables. Given this criterion, in the 2700-word test set 53% of all the words are strong-initial. The network performance, as shown in Fig. 9, is proportionally skewed toward successful detection of strong-initial words to a striking degree ($\chi^2_{(1)} = 77.2$, $p < .001$).

A similar result was obtained when we changed the definition of weak syllables to just those involving /ə/ ($\chi^2_{(1)} = 70.4$, $p < .001$). A corollary of this behavior is that the model will segment more before open-class words, and examination of the totals of hits before open- as opposed to closed-class

[8] This point does not apply to a more constraint-based approach in which, for instance, segmentation information simply changes the activation levels of different lexical hypotheses, as in Norris's SHORTLIST model.

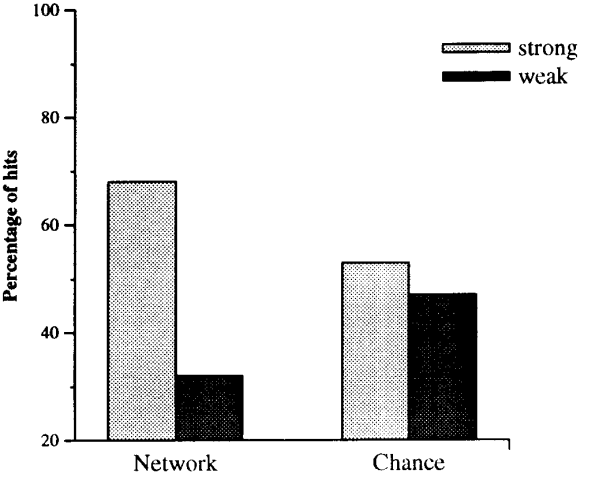

FIG. 9. Segmentation before strong and weak initial syllables.

shows that this is the case. Figure 10 shows that the beginnings of open-class words are much more likely to be detected than the beginnings of closed-class items ($\chi^2_{(1)} = 14.0$, $p < .001$). Note also that the boundaries with which the model has most difficulty are the closed–closed boundaries, thus strings of closed-class words such as *up to the* are less likely to be segmented than strings of open-class items.

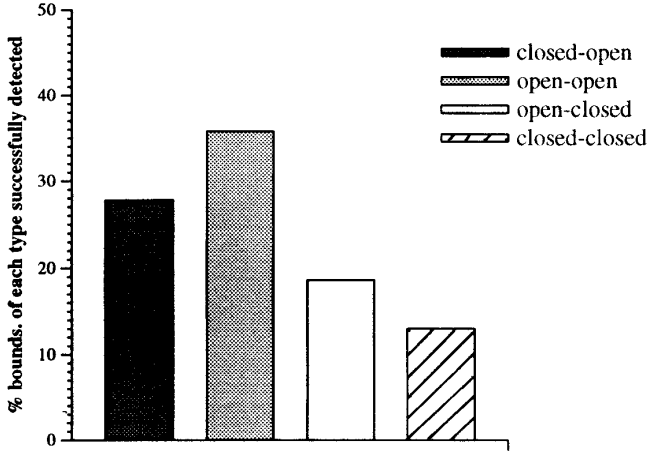

FIG. 10. Network performance in predicting open- and closed-class words, showing segmentation behavior across the different boundary types.

An inspection of the phones before and after which the model most frequently placed boundaries revealed that the choice of segmentation point was not simply predicted by the frequency of the previous and next segments: for instance, the model segmented much more after [z] than the frequency of this item would dictate, presumably due to plural endings. There was also some slight evidence that coarticulation provided cues, since the model tended to segment after certain co-articulated items (e.g., [k$^{\neg}$] − unreleased /k/) much more than it would by chance. However, remember that our coarticulation retranscription rules applied both inter- and intralexically, so we have not simply built in this cue. This effect is due to distribution of contexts where coarticulation can occur; these are more frequent interlexically.

Up to this point we have discussed only the individual boundary decisions that the network makes. We now consider the pairings of these individual segmentations—the words that emerge from the network. A word count of the LLC revealed that 40% of all tokens were open-class, so one would expect that this ratio would hold in network output, all other factors being equal. Although the network does not segment more whole words from the test stretch than it would by chance, of the correctly extracted tokens 59% are open-class. This is significantly more open-class items than one would expect from random segmentation performance: ($\chi^2_{(1)} = 19.46$, $p < .001$).

These results are interesting, as they suggest that the development of the MSS might not be due to the nature of the nuclear vowel, but might in fact emerge from predictions from the prior phonetic context. Our network model has no ''retrodictive'' capacity; it does not buffer the input and make a *post hoc* decision about a word boundary based on the nature of the following vowel. Our results suggest that a strong-syllable segmentation strategy could emerge from the difficulty of predicting the onsets of such syllables given their distributional statistics and prior context alone; in our model such information is the only possible source of the effect that we have demonstrated. Positions in which the model predicts a boundary are likely to be those in which there is a high degree of freedom in continuations for the string (high perplexity) and hence a low predictability. Therefore, as we might expect, closed-class words are more predicted by context and hence are less likely to be split up by the network.

In summary, the picture that emerges from these results is not in conflict with the MSS; rather, it provides a computational perspective on how the MSS could be acquired. An informationally primitive, purely bottom-up account can produce behavior which mimics the observational correlates of the MSS, but which does not rely on vowel quality, and hence has no need to rely on the *a priori* perceptual salience of strong syllables (cf. Cutler & Mehler, 1993). It is possible that some of the information used in this developmental model persists in the adult and is used for segmentation, even though more sophisticated knowledge about word boundary distribution may also be

available. Although the results of the network model are qualitatively similar to the MSS in some respects, the segmentations involving the closed-class words complement the MSS.

Adding Categorial Knowledge

The results presented in the previous section were obtained by segmenting with raw scores that were not normalized for phoneme type: thus, for instance, a particular error score was treated identically when predicting both the frequent segment /ə/ and the much less frequent segment /ʒ/. This procedure can be seen as simulating the phase of infant development in which phonemic categories, and information about their frequencies, are not yet available to the infant. However, we know that within the first year of life the child's phonological space is becoming structured into phonemic categories (see, e.g., Kuhl, 1983; Werker, 1993), and there is sufficient exposure to the language to provide accurate distributional statistics. We simulated the effect of this emergence of phonemic categories in our model to determine whether the qualitative pattern of segmentations would remain constant.

We carried out the same segmentation procedure as before, but this time normalizing for phoneme type by dividing each prediction error score by the frequency of the segment being predicted. We found an entirely different pattern of results with respect to strong syllables and word class, compared with the results described above: the network no longer mimicked the MSS. Segmentation before strong as opposed to weak syllables was not significantly different from chance: ($\chi^2_{(1)} = 0.387$, n.s.). Similarly, segmentation before open- as opposed to closed-class items was at chance levels: ($\chi^2_{(1)} = 0.035$, n.s.). Furthermore, using phoneme-normalized scores, 78% of correctly extracted word tokens were closed-class, in contrast to the 41% obtained previously with raw scores. This figure differs significantly from the expected distribution: ($\chi^2_{(1)} = 8.07$, $p < .005$).

Segmentation behavior changes in this way when scores are normalized because closed-class words, being very frequent in the language, contain the most frequent phonemes. Therefore, the network will predict these phonemes more often, and activate their representations more highly, compared with phonemes that do not occur often in closed-class words. Because predicting these segments is easier, error scores are lower. Hence normalizing for phoneme type will augment the error scores for phonemes that most often occur in closed-class words and effectively increase the probability of boundaries being proposed before such segments. This interpretation was borne out by an inspection of the phones that the model placed boundaries before and after most frequently. When we compared these phones with those produced before normalization, it was clear that the result of the normalization had been to boost the number of segmentations that occur before and after more frequent

segments. After normalization, the two sets of most frequent segments before and after boundaries intersected to a large extent.

In summary, once the experience of the processor in accruing frequency information about the speech stream is taken into account, segmentation behavior changes quite dramatically: the network does not now mirror the output of the MSS; rather the network detects the previously undersegmented closed-class boundaries more successfully than before. We will argue below that this limited window in which the generalization represented by the MSS is visible may be seen as a critical period effect.

GENERAL DISCUSSION

In general our approach has been somewhat different from that of the majority of connectionist models of language. Whereas many authors instantiate a particular psycholinguistic process in a working model, our models are not representations of psycholinguistic processes per se. Rather, we use statistical models as tools with which to encode phonological information. These models can then be used to assess the utility of such information in particular psycholinguistic processes. This approach has the principle advantage of permitting our modeling to be full scale; whereas a standard connectionist model may be forced to restrict its input in some way (e.g., small lexicon, only monosyllabic words, limited inventory of phones) we can use input that is truly representative of conversational speech.

In order to stand as a valid model of human behavior, and particularly acquisition, a statistical system must be derived from input that is representative of genuine natural language input, yet this fact seems to be forgotten in much connectionist modeling work. Linked to this issue is the tendency to use training data that are idealized and noise-free. One of the goals of this paper has been to show that real, noisy data can be used in psycholinguistic modeling and that such data allow our models to claim increased ecological validity.

The main empirical claim behind our approach is that subregularities within a domain can be, and are, exploited to the extent that they make useful predictions. In our case the subregularity is phonotactics, the sublexical distributional regularities of phonology. Any native speaker of English instantly knows that /strɪnk/ could be a word, while /nkstrɪ/ cannot, suggesting that we do have rapid and automatic access to information about legal English phonotactic structure. Evidence from work by Cowan (1991) has shown, albeit for nonword stimuli, that adults are sensitive to the frequency of repeated patterns in speech and will use that information in segmenting speech. There is evidence to show that such sublexical information can explain results that would otherwise require higher-level explanations. Foss and Gernsbacher (1983) demonstrated that vowel length could explain anomalous phoneme monitoring data; Chater, Shillcock, Cairns, and Levy (submitted for publica-

tion) showed that the data used by Elman and McClelland (1988) as evidence of top-down penetration of perceptual processes could in fact be given a bottom-up explanation based on phonotactic factors. Further, it seems that people are particularly good at extracting simple patterns, spatial or temporal, in input: we exploit redundancy, possibly for the purpose of efficient memorization. This is the interpretation that some researchers have placed on results from the artificial grammar learning paradigm, and in particular the phenomenon of transfer across domains (see Perruchet & Pacteau, 1990), in which subjects appear to respond to low-level regularities in pseudolinguistic input and do not internalize complete grammars. Sensitivity to simple patterns would seem to be present across various domains (e.g., vision, audition), and so the general model that we propose does not necessarily have to be initially part of a specialized language processing system. We hypothesize that subsegmental and/or segmental distributional information will be useful for segmentation in human languages other than English.

Although we believe that distributional information may be used by the human segmentation mechanism, this does not mean that other sources of information are neglected. We have seen that although phonotactics provides an initial purchase on the segmentation problem, it cannot be a complete account of adult segmentation behavior or acquisition. Rather we believe that phonotactics can be used as one of a set of conspiring bottom-up cues to possible boundaries, which mark input strings with boundary hypotheses (possibly weighted probabilistically). These cues will be both qualitative, such as acoustic juncture markers and metrical cues, and statistical, such as phonotactics. The collusion of such cues either will be sufficient to solve the segmentation problem or will need to be augmented by the introduction of higher-level knowledge.

We have presented, above, two different types of model: supervised and unsupervised. Both classes of model are bottom-up, either weakly or strongly so in the senses already defined. These two classes of model can be applied to adult segmentation behavior and development, respectively, as we now discuss.

The Role of Phonotactics in Explaining Adult Behavior

We have shown that probabilistic *n*-gram models can be extremely powerful word boundary detectors given a full phonemic transcription: the best performance for a trigram model was 93% of all boundaries detected with a hits:false alarm rate of 9:1. When this transcription was replaced with a much more robust broad-class transcription that should be available to listeners even in noisy circumstances, segmentation performance was degraded slightly: 74% detection at 1.5:1. This source of information is too predictive to be ignored by cognitive systems that in general have an aptitude for sensitizing to distributional information.

The supervised *n*-gram model was trained with labeled input and therefore is best seen as a model of adult competence: the learning in the model does not correspond to human development; only the resulting trained model has any ontological significance. Explicit training is not normally a part of human development, so we need to explain how the model can come to be part of the human language processor. Phonotactic information with word boundaries could be stored efficiently in a low-level autonomous process and could develop from the correlation of successful output activity in the different processor components—the standard method for learning in modular systems. However, it would be hard to find experimental psycholinguistic evidence that specifically supported the supervised *n*-gram model, since such evidence would be hard to dissociate from the explanation that phonotactic knowledge arises from rapid probability calculation over lexical phonological representations and not from a separately encoded phonotactic module. However, Cutler et al.'s (1992) demonstration of the exclusivity of segmentation algorithms in accomplished French/English bilinguals (any one individual appears to use either a metrical or a syllabic strategy) suggests that segmentation algorithms are not simply epiphenomenal on lexical access: a French/English bilingual using a syllabic strategy still has an English lexicon which could support a metrical strategy.

As we noted above, the use of bottom-up information is compatible with both interactive and modular cognitive infrastructure. However, showing that bottom-up information can account for a phenomenon such as segmentation strengthens the credibility of the modular position by virtue of Occam's razor: if segmentation can be carried out by making reference to some simple acoustic/prosodic/distributional cues, then why invoke the overhead of a lexical search?

However, we have shown that phonotactic information is at best a partial solution to the segmentation problem, especially if a full phonemic transcription cannot be taken for granted. We have ignored prosodic information in the studies reported above. Elsewhere (Shillcock, Chater, Cairns, & Levy, 1995), we have extended the distributional approach described here to the prosodic transcription in the LLC, showing the utility to lexical segmentation of an idealized prosodic transcription which extends over veridically demarcated syllables and tone units.[9]

[9] Three prosodic tracks were employed: contour (rise, fall level, null), stress, and boosters (sudden pitch shifts), the last two being single-valued. The statistics for the bigram transitions were calculated for this three-tiered transcription, and low probability transitions were assessed as lexical boundaries. The results, for instance a 50% hit rate with a 12.5:1 hits:false-alarms ratio, reflect the use of the veridical syllabic boundaries across which the prosody was marked discontinuously. Nevertheless these figures show the potential utility of prosodic information to segmentation.

Moreover, it is probable that a strictly left-to-right model is not an accurate description of human behavior, since as many as 20% of words in normal speech are recognized, if at all, only after their acoustic offsets and after information has arrived about the following word(s) (Bard, Shillcock, & Altmann, 1988). Thus a retrodictive element is probably necessary in any bottom-up account of segmentation; Content and Sternon (1994) provide a technique for making a segmentation device take account of left context. If it turns out that a bottom-up model is incapable of capturing all the data relevant to segmentation, then lexical information must be used in segmentation. However, the need for lexical information would still not necessitate an interactive architecture; instead lexical information can be integrated with stimulus information in the manner of the FLMP model of Massaro (1992). The crucial distinction between the FLMP approach and the interactive approach is that under the FLMP lower-level representations cannot be altered by higher-level information: thus modularity is preserved. An alternative would be to use bottom-up information to restrict candidate lexemes that are submitted to an interactive component, as in the Norris (1994) model.

Modeling Development with Bottom-up Phonotactic Models

The unsupervised *n*-gram and neural network models, because they learn to segment without the influence of an external teacher, can be seen as bootstrapping models of infant development. We have provided a computational underpinning to the claim that low-level phonotactics could be used by a neonate as a cue for breaking up the continuous stream of input speech. We have argued that while *n*-grams provide a good basis for such a model, their inherently categorial nature renders them incapable of addressing the issues that arise in the first, precategorial, stages of development. Therefore, we used a feature-based neural network model which attained a reasonable level of segmentation performance at low detection rates (taking pausing into account, 32% of boundaries with a hits:false-alarms ratio of 2.4:1). Higher detection rates were accompanied by a sizable proportion of false-alarms, suggesting that the optimum compromise between hits and false-alarms should reflect the relative problems posed by under- and oversegmentation. Undersegmentation certainly occurs in children's speech processing, is probably a relatively temporary problem as two- and three-word strings that have been lexicalized are unlikely to be reinforced by repeated exposure, and may possibly be subsequently reanalyzed into their constituent parts. Oversegmentation appears to be less desirable than undersegmentation, having only the potential saving grace of sometimes revealing morphological structure. However, the presence of resyllabification across word boundaries may mean that erroneous segmentation cues (in this case prosodic) are a permanent feature of speech processing, even for the adult.

We have provided an account of how the MSS could arise without recourse

to positing metrical information as part of a genetic endowment. The network segmentation performance was significantly biased in favor of detecting open-class words that have strong initial syllables. Furthermore, we have shown that our model's mirroring of the MSS disappears when we add knowledge about the frequencies of individual phoneme categories: the open-/closed-class advantage in segmentation is reversed and detection of closed-class words becomes favored. Therefore, phonotactics could provide initial segmentations from which the utility of the MSS could be recognized in infancy. Once the MSS is in place, and the infant's phonological space comes to be securely structured with the phonemic categories of English, then phonotactics could still contribute to segmentation by isolating the closed-class items, which are typically problematic for the MSS. Categorical perception of phonemes emerges in the first year (Kuhl, 1983) and develops thereafter (Best, McRoberts, & Sithole, 1988; Werker & Tees, 1984; Werker, 1993). Our results suggest that there is an early perceptual disadvantage in the processing of closed-class words in continuous speech if phonemic categories have not yet been established.[10]

The disappearance of MSS-like behavior in our model when segmental frequency statistics are taken into account raises the possibility of a critical period for realization of the MSS. If we assume that the categorial knowledge necessary to compile such frequency statistics is pervasive after a certain stage of development, then phonotactic information will be able to bootstrap the MSS only in the precategorial phase. Whereas standard accounts of critical period phenomena suppose that the driving force behind the changes is the infant's maturation (e.g., Lenneberg, 1967), our work suggests that there may be alternative explanations arising from informational constraints and interactions external to the child, but intricately linked to the learning process. Cutler et al. (1992) show that English/French bilinguals divide into two groups, one that uses the MSS and one that does not. They suggest that this division may be the result of the predominance of one or other language in the very early linguistic environment of the child. Our methods provide one possible computational explanation of this result.

So, phonotactics is a possible source of reliable boundary finding information, but is there any evidence that infants actually use this strategy? The extrapolation of patterns of data in sensory input is one of the most fundamental and sophisticated behaviors that infants exhibit. Given that the child is innately predisposed to extract linguistic and other regularities from input, and to store speech input, it would in fact be surprising if any information

[10] Note that this period, the very earliest months of infancy, precedes the period addressed by existing demonstrations of differential processing of closed- and open-class words in English (e.g., Gerken & McIntosh, 1993).

source that was as informative as the phonotactics of English was ignored. Furthermore, experimental evidence shows that infants *are* sensitive to the sequential constraints of natural language (see Bertoncini & Mehler, 1981; Friederici & Wessels, 1993; Jusczyk, Friederici et al., 1993). Even so, if we assume that the child is sensitive to such regularities, then we must answer the most difficult question: how does the child come to know that these patterns predict the boundaries of chunks of meaning?

Our conclusions do not necessarily rely on the use of prediction in our simulations. The same effect can be obtained with the use of short-term storage. In the case of our connectionist model, the task is to predict the next segment, and when this task is difficult, it is more likely that a word is starting. So, if an infant were constantly trying to predict input, dissonance between predictions and reality would provide the necessary impulse for our segmentation strategy to emerge. However, another possible instantiation of the strategy is to remember stretches of input—difficulty of encoding is proportional to the probability that the model is trying to remember over a word boundary. If we assume that low-level phonological memory is inherently imperfect and noisy, then it is easy to see how knowledge of phonotactic structure might be used to smooth over the effects of some of the noise by making the past more predictable. This is the instantiation of our abstract model in the mind which we favor until more experimental evidence becomes available.

CONCLUSIONS

In conclusion, we have shown that the phonotactics of spoken English contain sufficient information for a completely bottom-up processor to be able to detect some one-third of word boundaries, when pausing is also taken into account, at a hits:false-alarms ratio of 2.4:1. Although this performance is modest of itself, it does represent a substantial initial purchase on the segmentation problem. We have argued that the undersegmentation represented by this result is not a severe problem for the claim that phonotactics has an important role in acquisition, first because such multiword strings do appear in child speech and can often be assigned meaning, second because such entries are relatively unlikely to be reinforced by repeated exposure, and third because these lexical items are presumably available for later reanalysis. Further, we have shown that such lexical boundary information is coincidental with syllabic boundary information for spoken English, underlining the rather ambiguous role of the syllable in English speech processing.

We have confirmed the utility of the Metrical Segmentation Strategy and demonstrated a way in which the MSS might be discovered by a data-driven approach. We have shown that the early use of phonotactic information produces segmentation behavior that resembles the MSS, preferentially discovering open-class word boundaries, producing an outcome in which the

generalization represented by the MSS might be recognized and instantiated in the processor. The statistics we present confirm the move, apparent in the MSS-related research of Cutler, Norris, and colleagues away from the previous emphasis on the open-/closed-class distinction. In the statistics we report, the MSS correctly identifies half of all word boundaries, confirming its value as a source of segmentation information.

We have claimed that sensitivity to phonotactic information represents a viable part of a wider developmental model of speech perception. Some more general issues arise from the results we report. First, in both the *n*-gram studies and in the neural network studies, very local transitions provide the relevant information. In some cases bigrams even produced segmentation results superior to those of trigrams. In other cases, trigram statistics informed by knowledge of veridical lexical boundaries constitute virtually a complete solution to the segmentation problem. The information necessary to identify any one speech segment is typically spread across the immediately adjacent segments, as a result of coarticulation. Thus, bigrams and trigrams represent psychologically realistic windows for segment identification, as well as constituting the easiest statistics to be computed by the generally limited processing resources available in infancy. For further demonstration of the utility of bigrams, see, for instance, the exploration of context-sensitive coding in word recognition by Marcus (1981, 1985), or the role of phonotactic range in predicting consonant acquisition in English (Shillcock & Westermann, 1996). The past two decades of psycholinguistic research have made it increasingly clear that the human speech processor operates very rapidly, integrating new phonological information into the interpretation of the speech virtually as soon as it becomes perceivable. Segmentation decisions carried out over bigrams represent the fastest, most local segmentation decisions it is possible to take; other segmentation strategies, employing vowel information for instance, will typically involve longer stretches of the speech stream. We might therefore expect the processor to employ the phonotactic statistics we have explored, in order to take advantage of the most incremental segmentation strategy possible.

We have demonstrated that normalizing for segment frequency dramatically changes segmentation behavior from a pattern of segmentations that resembles the MSS to one that does not. Normalizing for segment frequency improves segmentation behavior, but the information necessary to normalize the calculations in this way can be accrued only after experience with the language. We therefore see a limited period within which the generalization represented by the MSS may be recognized. In effect, the normal process of accumulating experience with the language gives rise to a critical period effect. Note that the utility of the MSS would not necessarily become apparent in this manner if English was being learned as a second language and the input was classified in terms of segment frequencies associated with the first language.

Finally, having demonstrated the potential utility of distributional statistics in the development of speech segmentation, we leave open the question of the extent to which such relatively primitive strategies survive in the repertoire of adult segmentation abilities. Many sources of information potentially contribute to speech segmentation, and each may have differing strengths when required to segment continuous speech at the beginning of an utterance, speech in poor listening conditions, speech containing new lexemes, rapid speech, and so on. Distributional statistics have a definite contribution to make in the differing demands of speech segmentation, both in adult comprehension and in language acquisition.

REFERENCES

Abu-Bakar, M., & Chater, N. (1993). Processing time-warped sequences using recurrent neural networks: Modelling rate-dependent factors in speech perception. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 191–197). Hillsdale, NJ: Erlbaum.

Bard, E. G., Shillcock, R. C., & Altmann, G. T. M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception and Psychophysics,* **44**(5), 395–408.

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behaviour and Development,* **4,** 247–260.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance,* **14**(3), 345–360.

Bhattacharjya, A. K., & Roysam, B. (1994). Joint solution of low, intermediate and high-level vision tasks by global optimization: Application to computer vision at low SNR. In M. Mozer, P. Smolensky, D. Touretzky, J. Elman, & A. Weigend (Eds.), *Proceedings of the 1993 Connectionist Models Summer School* (pp. 39–47). Hillsdale, NJ: Erlbaum.

Brent, M. (1993). Minimal generative explanations: A middle ground between neurons and triggers. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 2–36). Hillsdale, NJ: Erlbaum.

Brent, M. R. & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition,* **61,** 93–125.

Cartwright, T. A., & Brent, M. R. (1994). Segmenting speech without a lexicon: Evidence for a bootstrapping model of lexical acquisition. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society* (pp. 148–152). Hillsdale, NJ: Erlbaum.

Charniak, E. (1993). *Statistical language learning.* Cambridge, MA: MIT Press.

Chater, N., & Conkey, P. (1993). Sequence processing with recurrent neural networks. In G. D. A. Brown & M. Oaksford (Eds.), *Neurodynamics and psychology* (pp. 269–294). London: Academic Press.

Chater, N., & Conkey, P. (1992). Finding linguistic structure with recurrent neural networks. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 402–407). Hillsdale, NJ: Erlbaum.

Chater, N., Shillcock, R., Cairns, P., & Levy, J. *Bottom-up explanation of phoneme restoration.* Manuscript submitted for publication.

Church, K. W. (1987). Phonological parsing and lexical retrieval. *Cognition,* **25,** 53–69.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation,* **1,** 372–381.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 133–163). Hillsdale, NJ: Erlbaum.

Cole, R. A. (1980). *Perception and production of fluent speech.* Hillsdale, NJ: Erlbaum.

Content, A., & Sternon, P. (1994). Modelling retroactive context effects in spoken word recognition with a simple recurrent network. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 207–212). Hillsdale, NJ: Erlbaum.

Cowan, N. (1991). Recurrent speech patterns as cues to the segmentation of multisyllabic sequences. *Acta Psychologica,* **77,** 121–135.

Cutler, A., & Mehler, J. (1993). The periodicity bias. *Journal of Phonetics,* **21,** 103–108.

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

Cutler, A. (1986). Auditory lexical access: Where do we start? In William Marsten-Wilson (Ed.), *Lexical Representation and Process,* (pp. 342–356), MIT Press, Cambridge, MA.

Cutler, A. (1993). Phonological cues to open- and closed-class words in the processing of spoken sentences. *Journal of Psycholinguistic Research,* **22**(2), 109–131.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation—Evidence from juncture misperception. *Journal of Memory and Language,* **31**(2), 218–236.

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

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

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology,* **24,** 381–410.

Doutriaux, A., & Zipser, D. (1991). Unsupervised discovery of speech segments using recurrent networks. In D. S. Touretzky, J. L. Elman, T. J. Sejnowski, & G. E. Hinton (Eds.), *Connectionist Models: Proceedings of the 1990 Summer School* (pp. 303–309). San Mateo, CA.

Elman, J., & Zipser, D. (1988). Learning the hidden structure of speech. *Journal of the Acoustical Society of America,* **83,** 1615–1626.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language,* **27,** 143–165.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science,* **14,** 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks and grammatical structure. *Machine Learning,* **7,** 195–225.

Finch, S., & Chater, N. (1993). Learning syntactic categories: A statistical approach. In G. D. A. Brown & M. Oaksford (Eds.), *Neurodynamics and psychology* (pp. 295–322). London: Academic Press.

Fodor, J. A. (1983). *Modularity of mind.* Cambridge, MA: MIT Press.

Foss, D. J., & Gernsbacher, M. A. (1983). Cracking the dual code: Toward a unitary model of phoneme identification. *Journal of Verbal Learning and Verbal Behavior,* **22,** 609–632.

Frauenfelder, U. H., & Peeters, G. (1990). Lexical segmentation in TRACE: An exercise in simulation. In G. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 50–86). Cambridge, MA: MIT Press.

Friederici, A. D., & Wessels, J. M. I. (1993). Phonotactic knowledge of word boundaries and its use in infant speech-perception. *Perception and Psychophysics,* **54**(3), 287–295.

Gaskell, G., & Marslen-Wilson, W. D. (1995). Modelling the perception of spoken words. In *Proceedings of the 17th Annual Meeting of the Cognitive Science Society* (pp. 19–24). Hillsdale, NJ: Erlbaum.

Gerken, L., & McIntosh, B. (1993). Interplay of function morphemes and prosody in early language. *Developmental Psychology,* **29,** 448–457.

Grosjean, F., & Gee, J. P. (1987). Prosodic structure and spoken word recognition. *Cognition,* **25**(1–2), 135–156.

Gupta, P., & Mozer, M. C. (1993). The nature and development of phonological representations: Network explorations. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Harrington, J., Watson, G., & Cooper, M. (1988). Word boundary identification from phoneme sequence constraints in automatic continuous speech recognition. In *Proceedings of the 12th International Conference on Computational Linguistics* (pp. 225–230). Hillsdale, NJ: Erlbaum.

Harris, J., & Lindsey, G. (1993). The elements of phonological representation. In Jacques Durand & Francis Katamba (Eds.), *New frontiers in phonology.* Harlow, Essex: Longman.

Harris, Z. S. (1955). From phoneme to morpheme. *Language,* **31,** 190–222.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence,* **40,** 185–234.

Huttenlocher, D. P., & Zue, V. W. (1983). Phonotactic and lexical constraints in speech recognition. In *Proceedings of the Third National Conference on Artificial Intelligence* (pp. 172–176).

Jusczyk, P. W. (1993a). How word recognition may evolve from infant speech perception capacities. In G. T. M. Altmann & R. Shillcock (Eds.), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting* (pp. 27–55). Hove, England: Erlbaum.

Jusczyk, P. W. (1993b). Discovering sound patterns in the native language. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 49–60). Hillsdale, NJ: Erlbaum.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development,* **64,** 657–687.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language,* **32,** 402–420.

Kaye, J. D., Lowenstamm, J., & Vergnaud, J. R. (1985). The internal structure of phonological elements: A theory of charm and government. *Phonology Yearbook,* **2,** 305–328.

Klatt, D. H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In *Perception and production of fluent speech.* Hillsdale, NJ: Erlbaum.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behaviour and Development,* **6,** 263–285.

Lehiste, I. (1971). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America,* **51**(6 (2)), 2018–2024.

Lenneberg, E. (1967). *Biological foundations of language.* New York: Wiley.

Lightfoot, D. (1991). The child's trigger experience—Degree-0 learnability. *Behavioural and Brain Sciences,* **14**(2), 364.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language,* **12,** 271–296.

MacWhinney, B. (1978). The acquisition of morphophonology. *Monographs of the Society for Research in Child Development,* **43,** 174.

Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology,* **10,** 29–63.

Maskara, A., & Noetzel, A. (1992). Forced simple recurrent neural networks and grammatical inference. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 420–427). Hillsdale, NJ: Erlbaum.

Massaro, D. (1994). Modularity of information, not processing. *Cahiers de Psychologie Cognitive,* **13**(1), 97–102.

Massaro, D. W. (1992). Connectionist models of speech perception. In *Connectionist approaches to natural language processing.* Hove, England: Erlbaum.

McClelland, J. L., & Elman, J. L. (1986). Interactive processes in speech perception: The TRACE model. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models* (pp. 58–121). Cambridge, MA: MIT Press.

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **20**(3), 621–638.

Marcus, S. M. (1981). ERIS—Context sensitive coding in speech perception. *Journal of Phonetics,* **9,** 197–220.

Marcus, S. M. (1985). Associative models and the time course of speech. *Bibliotheca Phonetica,* **12,** 36–52.

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

Mehler, J., Dupoux, E., & Segui, J. (1990). Constraining models of lexical access: The onset of word recognition. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 236–262). Cambridge, MA: MIT Press.

Menn, L., & Matthei, E. (1992). The ''two-lexicon'' account of child phonology: Looking back, looking ahead. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 211–248). Timonium, MD: York Press.

Norris, D. G. (1990). A dynamic-net model of human speech recognition. In G. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and cognitive perspectives.* Cambridge, MA: MIT Press.

Norris, D. G. (1993). Bottom-up connectionist models of ''interaction.'' In G. Altmann & R. Shillcock (Eds.), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting.* Hillsdale, NJ: Erlbaum.

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

Norris, D. (1992). Connectionism: A new breed of bottom-up model? In *Connectionist approaches to natural language processing.* Hillsdale, NJ: Erlbaum.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **21**(5), 1209–1228.

Norris, D. G., McQueen, J., Cutler, A., & Butterfield, S. (1995). *The possible-word constraint in the segmentation of continuous speech.* Unpublished manuscript.

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

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General,* **119,** 264–275.

Peters, A. M. (1983). *The units of language acquisition.* Cambridge: Cambridge Univ. Press.

Redlich, A. N. (1993). Redundancy reduction as a strategy for unsupervised learning. *Neural Computation,* **5**(2), 289–304.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

Shillcock, R., Bard, E. G., & Spensley, F. (1988). Some prosodic effects on human word recognition in continuous speech. *Proceedings of Speech '88, 7th FASE Symposium* (pp. 827–834).

Shillcock, R. C., Chater, N., Cairns, P., & Levy, J. P. (1995). *Speech segmentation: A paradigm*

example of constructivist language acquisition. Poster presented at *Architectures and Mechanisms for Language Processing '95,* Edinburgh.

Shillcock, R. C. & Westermann, G. (1996). The role of phonotactic range in the order of acquisition of English consonants. Poster presented at ICPLA '96, Munich.

Shillcock, R. C., Hicks, J., Cairns, P., Levy, J., & Chater, N. (in press). A statistical analysis of an idealised phonological transcription of the London–Lund corpus. *Journal of Computer Speech and Language.*

Shillcock, R., Lindsey, G., Levy, J., & Chater, N. (1992). A phonologically motivated input representation for the modelling of auditory word perception in continuous speech. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 408–413). Hillsdale, NJ: Erlbaum.

Shillcock, R. C., Levy, J., & Chater, N. (1991). A connectionist model of word perception in continuous speech. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp. 340–345). Hillsdale, NJ: Erlbaum.

St John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence,* **46**(1–2), 217–257.

Suomi, K. (1993). An outline of a developmental model of adult phonological organization and behaviour. *Journal of Phonetics,* **21,** 29–60.

Svartvik, J., & Quirk, R. (1980). *A corpus of English conversation.* Lund: LiberLaromedel Lund.

Tanenhaus, M. K., & Lucas, M. (1987). Context effects in lexical processing. In U. Frauenfelder & L. Tyler (Eds.), *Spoken word recognition.* Cambridge, MA: MIT Press.

Vroomen, J., & van den Bosch, A. (1992). *Speech segmentation in a connectionist model.* [Unpublished manuscript]

Werker, J., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development,* **7,** 49–63.

Werker, J. F. (1993). Developmental changes in cross-language speech perception: Implications for cognitive models of speech processing. In G. T. M. Altmann & R. C. Shillcock (Eds.), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting.* Hove, England: Erlbaum.

White, H. (1992). *Artificial neural networks.* Oxford: Basil Blackwell.

Williams, G., & Brockhaus, W. (1992). Automatic speech recognition: A principle-based approach. In A. Göksel & E. Parker (Eds.), *Working papers in linguistics and phonetics* (Vol. 2, pp. 371–401). London: School of Oriental and African Studies.

Wolff, J. G. (1977). The discovery of segmentation in Natural Language. *British Journal of Psychology,* **68,** 97–106.

(Accepted August 1996)