Chapter 11

Constraining Models of Lexical Access: The Onset of Word Recognition

Jacques Mehler, Emmanuel Dupoux, and Juan Segui

Years of excellent research at many distinguished institutions may license the conclusion that continuous speech has no obvious cues that facilitate the segmentation processes. However, even if cues indicating boundaries or natural segments have not been found, it would be rash to claim that they cannot be found. The lexicon is acquired, and infants are provided with little information about words pronounced in isolation. They are mostly stimulated by continuous speech. Thus, if word boundaries were not available, the acquisition of a lexicon would be incredibly difficult, if not outright impossible. Unfortunately, studies in this domain are by and large not available. So let us take a step back and ask what is known about how adults access lexical information.

It has classically been assumed that the segmentation of continuous speech into words is made possible by using the lexicon. Models of speech processing that maintain that perception is contingent on lexical access favor the view that the signal is matched against the lexicon without any prior segmentation. Segmentation is viewed in these models as a by-product of lexical access. We correctly parse the sentence "The book is on the table" because *bookiso* is not a word. However, these kinds of procedures rapidly run into difficulties. For instance, barring allophonic variations, it may be argued that the phonological realization of the orthographic strings *booki* and *bookis* can be interpreted as words. Moreover, these solutions are not appropriate for coping with frequent instances of phonetic and phonological ambiguity in natural languages, e.g.,

Good candy came anyways.
Good can decay many ways.

Faced with such potential ambiguities, automatic recognition systems adopt two solutions. The first one is to have speakers segment their own speech by inserting pauses between words and/or syllables before the

information is fed to the computer (Bahal et al. 1981). The other solution is to rely heavily on constraints derived from higher levels of processing to disambiguate the various segmentations of words embedded in sentences. Unfortunately, such a solution is not sufficient, as attested by the relatively poor score obtained by recognition systems in continuous speech. An even more severe problem for the first solution is that the proposed routines cannot be used by the young infant who is in the process of acquiring spoken language. For one thing, the infant lacks a lexicon, a syntax, and a semantic and pragmatic system. Yet children have to solve the segmentation problem to acquire the lexicon, and they cannot do this in a top-down fashion. This poses a paradox: if infants have to extract words in order to construct their lexicon, how can this process possibly rely on the lexicon? We are apparently back to square one with what seems an unsolvable problem.

In this chapter we will argue that the problems of segmentation might be lifted if we posit a *prelexical unit* available to infants for the construction of lexical entires and the extraction of the phonological structure of speech sounds. We will also provide evidence that infants use a unit that roughly corresponds to the syllable. We suggest that the same type of processing unit is used for speech acquisition in infants and speech recognition in adults. This proposal is compatible with recent data supporting the role of the syllable in adult speech processing (see Segui, Dupoux, and Mehler, this volume).

SARAH, a framework presented in more detail at the end of this chapter, proposes that a syllable-size unit is necessary for the child to correctly parse continuous speech. Each natural language provides cues that may be used by a speech-processing device. Indeed, we argue that continuous speech incorporates acoustic and prosodic cues that facilitate the identification of linguistically relevant boundaries. Such cues ought to be particularly useful during speech acquisition.

Finally, we address the issue that the details of processing and the specification of linguistic units is a function of linguistic experience (Abramson and Lisker 1965, Cutler et al. 1986). Recent results suggest that infants use less-specific processing units than adults (Trehub 1976; Werker and Tees 1984; Best, McRoberts, and Nomathemba 1988). We present some suggestions concerning the mapping of a *universal* system onto a *specialized* one.

We believe that a good model of speech perception must not only incorporate an explanation of lexical access but also present a credible account of how the organism compiles relevant linguistic information to

acquire a lexicon. In view of what we do know about the input available to infants, a speech perception routine that in *principle* cannot be acquired remains an inadequate psychological model. Nonetheless, most models of speech perception pay little attention to such learning constraints. In contrast, we posit a rather close *correspondence* between the structures or modules used by young infants and adults and argue that the data collected at both ages can be used to constrain models of speech processing at the initial and stable states.

Processing Units before the Acquisition of the Lexicon

A variety of computational and psychological models have been proposed for spoken-word recognition in adults. Each of these requires the specification of a code in which lexical entries are expressed. Such a code corresponds to the the basic psychological or computational unit in terms of which the acoustic wave is analyzed. After many decades of research, no consensus has been reached concerning its general properties.

An important factor that distinguishes current models in speech perception from each other concerns the *linguistic status* of the processing units. In most computational models, processing units have no clear linguistic status: vectors of linear-predictive-coding (LPC) coefficients, spectral templates, formant trajectories, etc. Klatt (1989) claims that this is desirable, since premature linguistic categorizations may induce virtually irreparable errors. In such models the only relevant linguistic unit is the word itself.

In contrast, in many of the psychologically motivated models, words are represented in terms of such linguistic units as features, phonemes, syllables, morphemes, etc. What differentiates these models is the degree of linguistic reality postulated. Models that acknowledge *strong linguistic reality* postulate processing units that are isomorphic to linguistic constructs. For instance, TRACE postulates phonemic units, and phonemes are recognized via the activation of distinctive features. Likewise, Treiman (1983) postulates the syllables as a psychologically real unit and proposes that the *internal structure* of the syllable (onset, coda, rhyme) as described by metrical phonology (Halle and Vergnaud 1987) is used in processing.

Other models adopt units that have *weak linguistic reality*. The units are usually also *similar* to the ones used by linguists, namely, syllables, morphemes, words. However, no a priori commitment is made to the internal *structure* of the units as postulated by linguistic theory. Empirical investigations seek to determine the nature of processes underlying the activation and are using the units to settle this point. Meanwhile, speech-

recognition models are being developed with only adults in mind. How well can such models account for speech perception in young infants before they acquire language?

The first problem that a child has to solve is how to identify auditory inputs relevant to maternal language. Clearly, models of speech acquisition must explain why the child does not attempt to construe every possible sound as a potential lexical item. Otherwise, how could we explain why infants do not acquire a lexicon containing chirps, burps, dogs' barks, engine noises, and so forth in addition to words? Theories of speech acquisition must address this issue.

As a matter of fact, if infants represent potential words in terms of purely acoustic information (e.g., spectral templates of say 10 msec each, LPC coefficients, Fast Fourier Transform parameters, etc.), they would not be able to separate a bark from a syllable, say [ba]. Nor would they be able to class together speech sounds, on the one hand, and Schuman's fantasies on the other. However, children do construct a lexicon leaving out noises, music, etc. Although children can recognize a bell, a dog's bark, or a car, they usually cannot say which dog they heard barking or what make of car has gone by. The only lexical compilation that is spontaneously established corresponds to the one used in speech. Thus it is unlikely that infants rely exclusively on acoustic characterizations of speech sounds. In all likelihood humans represent speech sounds by means of structures that are linguistically specific, as is posited by models that postulate a strong or weak linguistic status for processing units in online speech perception. Thus there is some motivation to view speech processing as different from that of other acoustic signals.

One of the most important hypotheses about speech processing was advanced by Liberman and his colleagues (1967), who postulated a special mode to process speech. The speech mode is different from the one used for processing other acoustic stimuli. The existence of a speech mode has motivated a large number of investigations in the last two decades. Some of these investigations directly explored whether speech is special, while others went about it in an indirect fashion. Some investigators argue that the available data show that infants are attracted toward speech sounds and that this makes it possible to demonstrate differential processing of speech and nonspeech stimuli. Colombo and Bundy (1983) reported that four-month-olds show a strong preference for hearing voices over hearing noises. Moreover, this preference is stronger when the voice is presented on the right side. More recently, Bertoncini et al. (1989) showed that four-day-old infants react differently to speech and musical sounds. Infants

show right-ear superiority for dichotically presented syllables and a left-ear superiority for dichotically presented musical sounds. The interaction between ear and type of stimulus was significant. These results suggest that very young infants do not have to learn how to process speech and nonspeech sounds differently. Other studies on young infants corroborate these findings. Indeed, Best et al. (1988) and Segalowitz and Chapman (1980) found similar results with a different technique. The human brain is programmed to process speech differently from other acoustic stimuli.

Eimas et al. (1971) attempted to evaluate the ability of one-month-old infants to distinguish and categorize speech sounds. Eimas (1974, 1975) showed that infants categorize stop consonants, liquids, etc., like adults. In short, the infant is born with the ability to discriminate and later to categorize any potential phoneme in any language. Furthermore, human infants are very skillful when it comes to mastering language. Mills and Meluish (1974) reported that fourteen-week-old infants suck more intensely and longer when hearing the voice of their own mothers as opposed to that of a stranger. A similar result was reported by Mehler et al. (1978) with eight-week-olds and by DeCasper and Fifer (1980) with twelve-hour-old neonates.

In brief, infants process speech sounds differently than nonspeech sounds. Indeed, they recognize individual voices soon after birth. These observations are important for any model of speech acquisition. For instance, a proposal compatible with the behavior observed in neonates is that infants are able to *represent* and *memorize* speech sounds in terms of specific abstract linguistic units. It is in virtue of this fact that speech sounds are special. It would appear that classifying a dog's bark and linguistic utterances as different sounds is not a problem for babies. They are already equipped with specialized machinery for processing and representing speech sounds. Thus the data collected with babies favors what we have called the *linguistic status* of the early speech-perception units. However, it is rather difficult at this time to say whether the data favors a weak or a strong linguistic status.

Fine-grained versus coarse-grained models for capturing phonotactics

We assume that humans rely on the speech mode to process incoming speech utterances and to represent these in an abstract format or a linguistic code. What are the properties of such a code? Two types of models can be distinguished on the basis of the size of the proposed processing units. *Fine-grained models* postulate that transduction is smooth and relies on virtually continuous information.[1] During information uptake, central

lexical representations are gradually activated and rapidly deactivated. The process continues until a unique word candidate is isolated. In lexical access from spectra (LAFS), the basic unit of analysis consists of centisecond spectral templates (Klatt 1977, 1989). Recently Marslen-Wilson (1987) has proposed that cohorts feed on such a flow of spectral templates. These models implement an optimal recognition strategy. The lexical system receives new acoustic information on-line, which enables it to select the best word candidate as soon as possible. Fine-grained models, however, need an additional processor to account for the extraction of the phonetic structure of nonword stimuli. In Klatt's initial proposal, LAFS was complemented by the SCRIBER system, another decoding network, whose purpose was to compute the phonetic structure of utterances directly from spectral templates.

In contrast to fine-grained models, *coarse-grained models* postulate that the information flow between peripheral and central levels is discontinuous and relies on rather large information units. In these models a proper intermediate processing level (the prelexical level) accumulates incoming acoustic information before releasing it to high levels. These models do not implement an optimal recognition strategy for speech, because nothing happens before a critical amount of peripheral information has been processed.

A number of mixed models have been formulated to take advantage of the processing efficiency of fine-grained models while still postulating a prelexical level. In these *intermediate-grained models* the information grains are some fraction of syllables. In TRACE the smallest processing unit that contacts the lexicon is the phoneme (McClelland and Elman 1986).[2] Marslen-Wilson (1984) proposed that the reduction of the cohort takes place phoneme by phoneme. Likewise, Cutler and Norris (1979) claimed that words are recognized via phonemes in English. The intermediate-grained models proposed so far have a special feature: they do not need separate processing routes to cope with words and with nonwords. Indeed, the phonetic code of an utterance can be directly extracted from the prelexical levels.

There are many reasons to prefer coarse-grained models. Intermediate-grained models can be classed within the coarse-grained models because they share with them the assumption that processing is discontinuous and based on an intermediary level between the signal and the lexicon, namely, the prelexical level. In fine-grained models of speech perception, phonology and prosody are not explicitly represented, at least at the front end of the system. This deficiency causes fine-grained models to be too powerful, since

they can in principle represent very unsystematic or unnatural sequences, e.g., sequences corresponding to a mixture of Serbo-Croatian, Chinese, and English pronunciations. Coarser representations like syllabic trees, morphemes, or word boundaries are needed. Of course, within a fine-grained framework it is conceivable that phonology (syllabic structure, metrical representation, etc.) is derived directly from spectra by a specialized device analogous to the SCRIBER system proposed by Klatt to compute phonological representations directly from the signal.

However, such models should also specify how phonotactic information interacts with the acquisition of a lexicon. One possibility would be that phonotactic information triggers a flag to signal whether the sound sequence is or is not a legal sequence in the language. For instance, such a flag would signal that *pst* is an interjection, *clapitre* a possible French word, and *dlavotnik* a foreign word. Still, this model is not parsimonious, since many of the phonological regularities have to be acquired again in the lexicon. Another possibility would be to postulate that phonological regularities could be compiled at a level halfway between the lexicon and the signal. Of course, such a proposal maps onto a coarse-grained model.

Coarse-grained models propose a level of representation that can potentially filter out sequences of sounds that are illegal in any language, e.g., *pst* and *dztlkfx*. Phonological representations rest on primitives like the syllabic structure and stress pattern of each word. Attempts to model the phonological representations of languages with systems that do not comprehend rules has been tried, but none has paid off (see Pinker and Prince 1988 and Lachter and Bever 1988).

In fact, the smallest segment that can be a word is the syllable. Thus infants, we suspect, compile a bank of syllables, or *syllable analyzers*. Each syllable is represented as the prototype value of coarse-grain sequences that correspond to a compatible phonetic transcript in the language. As soon as the bank of coarse-grained detectors is compiled, lexical acquisition becomes a matter of storage and retrieval. Indeed, metrical and phonotactic regularities can be used to constrain the extraction of word boundaries and provide cues to the morphological components of the system. From the acquisition point of view a coarse-grained model is thus more plausible than a fine-grained model. We are nonetheless aware that this sketch leaves many problems unsolved.

What are the linguistic levels to which the infant pays particular attention? Are infants sensitive to both coarse- and fine-grained speech information, or do they focus preferentially on one of these? Mehler et al. (1978) showed that babies' differential activation to their mothers' voices

is not observed when normal intonation is disrupted, which suggests that *prosodic contour* is an important component of infants' speech processing. This finding is compatible with Fernald and Kuhl's (1987) claim that infants prefer to listen to "motherese" rather than to normal speech passages.

It would be rather unfair, however, to suggest that infants are exclusively limited to processing global parameters. In fact, as we showed before, infants are also excellent at discriminating such minimal speech contrasts as [pa] versus [ba], [ra] versus [la], and [pa] versus [ta]. Moreover, infants discriminate very short CV syllabic onsets (spliced from full syllables) where the consonant or the vowel of the full syllable from which they were spliced differs (Bertoncini et al. 1987). This behavior meshes well with adult perceptions. Indeed, adults claim that the very short stimuli give rise to the same phonological representations as the full syllables from which they were derived. Though the behavior of the infants is compatible with adult perceptions, more studies are necessary before it can be argued that infants categorize these short stimuli in the same way as adults. Since infants are so good at processing speech, the next issue that arises concerns the unit or units used to represent and memorize segments of speech. None of the results reviewed so far bears on this issue. Fortunately, a few results can be mentioned in this context.

Bertoncini and Mehler (1981) showed that neonates find it difficult to discriminate between synthetic [pst] and [tsp]. Yet young infants have no difficulty in discriminating [pat] from [tap]. Notice that the segments that indicate the difference between tokens is the same in the two cases, namely, a difference in serial order. Piaget and his collaborators have demonstrated that infants and very young children are not very good at dealing with serial order. How come, then, infants distinguish [pat] from [tap] but not [pst] from [tsp]? In the [pst]-[tsp] case the signals are not well- formed, while in the [pat]-[tap] case they are. However, when we add a vocalic context to the exact same nonspeech segments, such as [upstu] and /utspu/, infants again discriminate the contrast with great ease. These results suggest that infants organize speech sounds in terms of syllables. A contrast embodied in a syllabic context is discriminated even though the same contrast is neglected in other contexts.

Neonates tend to represent syllables as rather global conglomerates. As they grow older, they tend to elaborate an increasingly refined representation. Indeed, in a series of experiments, Jusczyk and Derrah (1987) and Bertoncini et al. (1988) showed that by two months, infants notice the addition of a new syllable into a corpus of four syllables used during

habituation, regardless of whether the new syllable's consonant, vowel, or both differ from the most similar habituation syllable. Furthermore, four-day-old infants notice the presence of a new syllable if it has at least one vowel different from that in the preceding syllables. In summary, these results indicate that the infant is fully capable of organizing speech signals in terms of syllables and prosody, i.e., coarse-grained units. Whether the speech signals are always organized in terms of syllables is an open question. However, we know that at least the ability to do so exists at birth.

Temporal normalization

A major problem that must be met by all models of speech recognition like the ones presented above is the fact that speech segments remain perceptually invariant, according to adults, over considerable changes in rate. Lenneberg (1967) claims that English speakers normally talk at a rate of 210 to 220 syllables per minute. He acknowledges, however, that rates of 500 syllables per minute are easy to attain for any normal speaker. Thus, mean syllable length may vary roughly from 300 msec down to 100 msec without affecting the ease or accuracy of comprehension. Chodorow (1979) and King and Behnke (1989) have shown that speech rates can be increased by as much as 60 percent without disrupting the identification of lexical items. At such rates, mean syllable duration is reduced to less than half of the original duration without changing intelligibility. Thus language users show recognition constancy for syllables and words pronounced at very different rates. Such an achievement is comparable to the perceptual constancies illustrated in the visual domain in most textbooks. As for other perceptual constancies, we have to ask whether speech recognition becomes independent of rate by learning or whether it makes language learning possible in the first place. To the best of our knowledge, all experiments that have assessed the intelligibility of compressed speech have used adult subjects.

It is difficult to say whether pre-lexical subjects recognize a syllable as invariant under many different durations. Yet there are informal indications, that infants identify syllables like *dog*, *mummy*, and *ball* regardless of speaker and durations. Parents never have to repeat a word uttered before so as to match its duration exactly with that of the prior pronunciation. How can they achieve such a performance? The accounts that provide a ready answer assume that the items are already represented in the lexicon and that the signal is recognized by some networklike routine. However, infants do not have a network with lexical items. They have to construct it.

Models of adult speech perception have proposed processing units with two types of temporal characteristics. The *durational* models claim that the minimal unit of analysis spans over a fixed temporal slice (Tyler and Wessels 1983, Marslen-Wilson and Tyler 1980, Salasoo and Pisoni 1985, Tyler 1984). Many computational models propose a purely durational processing routine, e.g., sampling the spectrum of the signal every 10 msec. *Structural* models postulate that the processing unit is constant under varying speech rates. Thus models that specify the primary inputs of their systems in terms of phonemes or distinctive features are synchronized with the rate of speech and qualify as structural models. This is also true of models that propose syllablelike units (Mehler 1981). We will see below that the structural status of the processing unit is crucial for solving the problem of normalizing the speech rate.

Temporal normalization, or freeing the categorization routines from duration, should be very difficult to learn. Durational models posit units of fixed duration, regardless of speech rate. Consequently, a change in speech should become highly disruptive for constancy extraction. To illustrate our point, let us consider the proposition that the first 150 msec of the signal are the processing unit for lexical access (Salasoo and Pisoni 1985). The contents of these 150 msec can, of course, vary dramatically when the speech rate changes. For example, one syllable in the word *capitain* in normal speech becomes over two syllables with a compression rate of 50 percent. If these 150 msec really function as processing units (units that have no internal structure and merely act as a code for accessing the lexicon), this implies that the child should build two access codes for the word *capitain*: *ca-pitain* and *capi-tain*, and for that matter, one for each possible speech rate. This solution, however, lacks design efficiency and raises serious acquisition problems. Indeed, it is mind-boggling to imagine how such a multitude of access codes can be learned and made to correspond to the same word without prior knowledge that they all represent the same segment of speech at different rates. Thus the absence of a normalization procedure seems to be fatal to strict-durational coarse-grained models.[3]

But how about a fine-grained model? Here the information content of a unit (e.g., a centisecond spectral template) seems relatively unaffected by variations in rate. However, speech compression modifies the time course of the activation of these units. If the speech rate is multiplied by two, the number of spectral frames that constitute each word is divided by two. This raises an important problem, since a given word is represented by a great many possible sequences of fine-grained units. If the lexical entries

are multiplied by the speech rates an important acquisition problem arises. Indeed, the child must identify all the lexical entries at all possible speech rates to have a lexicon. But to do this the child must identify words spoken at different rates as tokens of identical types. Although we have no idea as to whether this is feasible or not, we are skeptical. Another possibility, proposed by Klatt in LAFS, is to introduce *self loops* into the decoding network. This amounts to introducing a technique of dynamic time warping directly into the network. Such a move turns a durational model into a structural model.[4] Indeed, how many loops will the network have to complete before it halts? Which cues will be used to make it halt? Only structural cues will do a satisfactory job.

Many models of speech recognition define phonemes as a conglomerate of features. In this manner the temporal characteristics of phonemes can be established over a wide range of speech rates. Indeed, mapping articulation parameters might be implemented in a recognition algorithm. However, this is not sufficient, because in such models, duration cues are established as a parameter within phonemes, rather than between phonemes. But the temporal properties of adjacent segments are by no means negligible. Miller and Liberman (1979) have shown that the duration of the first formant transition that distinguishes [ba] from [wa] varies with the duration of the vowel [a]. Hence, an ambiguous sound somewhere between [ba] and [wa] is perceived as a rapid [wa] if the vowel is short and as a slow [ba] if the vowel is long. Moreover, there is some evidence that the average speech rate of the preceding sentence context influences perceptual boundaries, e.g., voice onset time (VOT) (Summerfield 1975) and vowel-duration cues (Port 1976, Port and Dalby 1982). This implies that durational parameters that span over many adjacent phonemes may play a crucial role in speech perception. Models that normalize for rate only in very short segments should induce the child to wrongly categorize sounds, which would make lexical acquisition very difficult, if not impossible. To account for speech normalization, it seems desirable to have a phonetic prelexical level where sequences contain information that spans several linguistic units.

The only possibility left within the two-dimensional continuum of speech-processing models appears to be a model with a rather large structural unit. This type of model accounts for the normalization problem in a rather neat fashion. For purposes of illustration, take a hypothetical syllablelike unit of processing, and let each syllable be represented in a bank of syllabic analyzers. During the processing of speech signals, dynamic time warping tries to adjust each syllabic frame for speech rate. Thus with

fast speakers, the spectral entries for each syllable analyzer are time-compressed by a given factor and compared to the signal. However, a compression or expansion factor should be constant within a given syllable. This constraint is absent in finer-grained models where each phonetic segment can be compressed or expanded independently. Thus coarse-grained models allow for normalization procedures that capture for each speech rate trading relations between the durations of consonants and vowels. Of course, contextual speech-rate effects have to be accommodated by postulating an extra mechanism, for instance, that during speech recognition, the bank of syllabic analyzers can come up with the best matching syllable as well as its duration. Thus, when listening to speech, one can arrive at the mean speech rate by averaging the duration of the syllables. This can, in turn, help to disambiguate such cases as [ba] and [wa]. For an ambiguous syllable, the bank of syllabic analyzers outputs two candidates: a rapid [wa] or a slow [ba]. A unique candidate can be selected by choosing which alternative best corresponds to the average speech rate of the context.

In summary and from the perspective of a speech acquisition, a coarse-grained structural model is desirable, since it allows for separate treatment of the problems of lexical acquisition and temporal normalization. Indeed, if the child has a coarse structural unit, his or her perceptual system can be tuned to factor out variations in speech rate. Once stable and invariant prelexical representations are available, the infant can rapidly compile a large lexicon without having to hear every existing word at every possible speech rate.

Everyday language incorporates important alterations in speech rate. Therefore, even very young children must have the capacity to extract constant categories in spite of major changes in speaking rates. What is the evidence that children perform temporal normalization? Miller and Eimas (1983) have shown that infants, like adults, tend to classify CV (consonant-vowel) syllables with a weighted function of formant transitions and the durations of the vowels, which suggests that infants' categorizations are determined by rate. In addition, infants classify sets of multisyllabic pseudowords by the number of syllables (Bijeljac-Babic, Bertoncini, and Mehler, in preparation). In the first experiment it was shown that infants react when changing from a list of bisyllabic items to a list of trisyllabic items or vice versa. Word durations were roughly matched in a control experiment, and the results stayed the same. A replication of these results would be interesting, since this work has im-

portant implications for understanding the relation between durational parameters and structural classifications in very young infants.

To sum up the results from research with young infants, it appears that these subjects tend to focus on coarse and global segments of speech rather than on fine ones. However, it must be acknowledged that there is no *direct* evidence that the young child uses syllablelike units for acquiring a lexicon or compiling prelexical representations. Obviously, the child has to rely on prelexical representations to acquire a lexicon. Thus children must engage in prelexical processing. The exact nature of the representations used and the processing on which young infants rely remains to be discovered. Psycholinguistic research provides some information that may be helpful in trying to establish the prelexical processes and units in children.

Cues to understanding infant and adult processing of language

The above arguments, although they remain inconclusive, bias us toward a model that is weakly linguistic, coarse-grained, and structural. Thus in order to steer a middle course, it would appear advisable to evaluate the syllable as a prelexical representation. Syllables are specific linguistic objects, they correspond to minimal speech gestures, and they are not to be confounded with other natural noises. Syllables are coarse-grained and can be used to compile the phonotactic regularity of the native language. Lastly, syllables are structural and are large enough to allow us to conceptualize algorithms for normalizing speech rates. Of course, we acknowlege that different languages may use different sets of coarse-grained, structural, weakly linguistic units (e.g., moras for Japanese). We will evaluate this issue in the next section.

We have argued on logical grounds that in speech acquisition infants probably use a coarse-grained, syllablelike unit, that is, a weakly linguistic, structural unit. Empirical data is scarce, but consistent with the existence of such a unit. How useful would such a unit be with the problem of segmentation and lexical acquisition?

There are several avenues open to explore the information that infants use in the course of lexical acquisition. In fact, there are many regularities at the prelexical level that the infant could use to discover boundaries between natural constituents of speech. Thus allophonic variations of a phoneme may signal the onset of words. For instance, in English the phoneme /t/ is always aspirated in word-initial position but not in intermediate or final positions. This generalization could be very useful to disambiguate sequences like *fast team* versus *fast steam*. Cutler (this volume) argues that main stress occurs mostly in word-initial position and

thus potentially signals word boundaries. A similar proposal can be found in Church 1987. However, for the infant to exploit such regularities, the notion of a word with its boundaries must come before.

In French, tonic accent is always in word-*final* position. Moreover, the distribution of syllables depends on their positions in a word. For instance, the syllables *-ique*, *-isme*, *-sion*, *-men* are very frequent in word-final position and quite rare in initial position. Such distributional properties may be useful. For instance, the words *measure* and *solution*, among many others, end with a syllable that is never realized in word-initial position.

Notice, however, that all these cues are language-specific and become available only after the child has established the proper processing routines for its own language. As was pointed out above, some cues even require the acquisition of the lexicon to become functional. How such language-specific cues are acquired is still an open question. But at least, seeing the segmentation problem from the point of view of speech acquisition obliges us to focus on previously ignored aspects of the linguistic signal. Infants, as we have seen above, show a remarkable capacity for processing speech at the prosodic level. Infants also have rudimentary notions of what counts as *phrase units* (see Hirsh-Pasek et al. 1987). We speculate that besides classifying utterances and identifying sentences and clauses, infants may also detect boundaries around words. For instance, when the child first hears a sentence like "Tommy, mange ta soupe" it will be parsed into clauses and then into syllabic templates. Syllabic templates are recognized by the bank of syllabic analyzers. The potential *word boundary detector* may use tonic accent to segment the sentence into potential words (*Tommy*, *mange*, *ta*, *soupe*). Thereafter the child will construct a lexical entry for each word. Consequently, the word *soupe* is represented by one syllable and the word *Tommy* by two. Notice the relationship of this proposal to Gleitman and Wanner's (1982) claim that children pay attention first to the stressed syllables in the sequence and only at a later age to the unstressed ones. This amounts to a segmentation routine related to the one we have in mind. "The child is prepared to believe that each wordline conceptual unit has an 'acoustically salient' and 'isolable' surface expression [that is] an abstract characterization of the sound wave whose surface manifestation in English is a *stressed syllable*" (p. 17).

Alternative mechanisms for surface acoustic cues have been conceived. For instance, the child may scan the continuous signal and simply store words if they are repeated. Words that occur often become represented and play a special role in lexical acquisition. Unfortunately, repetitions have to be found, and that is a major part of the problem that speech-

acquisition models have to explain. First the problem of word identification in continuous speech must be solved. Then maybe it will become possible to understand how the lexicon for speech recognition is compiled.

Like Gleitman and Wanner, we argue that the child requires surface acoustic cues to segment the speech stream. Furthermore, we speculate that such cues likely continue to be available to the adult. We are aware that hypothetical cues have been postulated to solve the child's bootstrapping problem but have not been uncovered by psychoacoustical research. This is, of course, a problem. However, there are many examples of cues that had not been psychoacoustically discovered but that we now know to be operational. Thus, although psychoacoustics has not yet determined the nature of the encoding, CV- and CVC-initial syllables are marked as such in the signal, as is shown by results obtained by Mehler, Segui, and Frauenfelder (1981). French subjects have shorter latencies to detect CV targets in words whose first syllable is CV and CVC targets in words whose first syllable is CVC. From the responses of the subjects, it appears that responses were elaborated before the lexicon had been accessed. This suggests that in French, initial syllables must be acoustically marked.

But what are the lessons that can be drawn from the infant data for *adult* models of speech perception? It is clear that the type of problem that the infant faces when it encounters its first utterance in English is not the same as that of recognizing one word out of a lexicon of more than 30,000 items. For adults, the problems of separating speech from nonspeech stimuli, extracting phonotactic regularities, and normalizing speech rates have already been solved. We have to acknowledge that in principle adults might use very different processing devices from those used by infants.

Yet current psycholinguistic investigations suggest that this is not the case. Cutler et al. (1983) claimed that the syllable is used as a prelexical representation by speakers of French (but not by speakers of English). Dupoux and Mehler (1990) provide evidence that the prelexical level relies upon large syllablelike units. These units are used either as codes for accessing the lexicon or to compute the corresponding underlying phonetic structure. The lexical code becomes available once a word is recognized. The prelexical code is used to recognize words and/or nonwords. Both codes share a common representation that is also the level at which the phonotactic generalizations of the language are captured. For more details on this issue, see Segui et al. in this volume and Dupoux and Mehler 1990.

The fact that the units used by adults and those used by infants are both syllablelike suggests that they are linked during development. If this is

true, much is to be gained in mutually constraining models by the joint study of adults and infants.

Some Open Issues on Language-Specific Tuning

Our proposal so far does not offer the solution to the child's bootstrapping problem. In fact, our proposal has only moved the bootstrapping problem from the lexical to the prelexical level. We posit that infants first parse continuous speech into discrete units and extract word boundaries. But the units and the word-boundary parameters in all likelihood vary from one language to the next. Each language has its own specificity. All languages use phonemes and distinctive features. However, all languages do not honor the same distinctions. For instance, some of them do not make a distinction between [r] and [l]; others use retroflex consonants, clicks, or tones. All languages have syllables, but some of them have syllabic reduction, ambisyllabicity, foots, or moras. All languages have words, but in some of them stress is in word-final position, in others it is in word-initial position, and in others stress can move and have contrastive value. During language acquisition, children learn the phonological and prosodic properties of their language, as revealed by the ability to classify a sequence as a possible word in the language or not. However, they have to do so before the acquisition of the lexicon.

These problems should be all the more difficult to solve if the child confronts several languages at the same time. However, infants raised in multilingual environments do not encounter any major problems in acquiring several parallel linguistic systems. They do not confuse phonology or prosodic properties across different languages. Obviously, adults easily discriminate the language they have mastered from a foreign one, but how can a child who has not yet mastered any language accomplish a similar *coup de maitre*? The child must have a way of segregating utterances from different languages. The set of criteria that might be available to them cannot be based on lexical representations, since the problem is precisely that of constructing separate lexicons for each language. It thus seems obvious that infants construct a representation of what counts as a permissible sequence in their language as opposed to other languages prior to lexical acquisition. There are many surface cues in terms of which languages might be distinguished, for instance, global prosodic contour, metrical structure, phonetic contrasts, syllable structure, etc. More recently Mehler et al. (1988) have shown that four-day-old infants react differentially to *utterances* in the parental language. Infants tend to suck

more when listening to French utterances if their parents speak French than with, say, Russian utterances spoken by a single bilingual speaker. Moreover, four-day-old infants respond differently when listening to a sequence of French sentences after a sequence of Russian sentences. Likewise, two-month-old American infants can also discriminate English from Italian. These results suggest that the human infant is able to classify novel utterances drawn from the parental natural language as tokens belonging to a type. They are not able, however, to discriminate tokens of one unfamiliar language from tokens of another unfamiliar language. Although these results need to be explored in further detail, we venture the prediction that it would be fairly difficult for a chimp or a chinchilla to perform like a human neonate even after a huge number of training sessions. A similar result was reported by Bahrick and Pickens (1988). These observations suggest that before they acquire a lexicon, infants can classify utterances into two categories: familiar and unfamiliar language. Mehler et al. (1988) showed that four-day-old infants are still capable of discriminating sequences of French and Russian after the signal is low-pass filtered, which masks most information with the exception of global prosodic contours.

In short, it seems that very early in life infants have a specialized device that classifies utterances of different languages. Furthermore, as we will see in the next section, within their first year of life, infants tune their perceptual system to the language in their environment.

Phonetic to phonemic convergence

The young infant's phonetic repertoire is far broader than that of adults. Monolingual adults are not very good at making discriminations on *phonetic* contrasts when these contrasts are not used in their language. Conversely, their performance is rather good for *phonemic* contrasts. A phoneme is the minimal difference in sound pattern that gives rise to different words. Thus, since /pet/ and /bet/ do not mean the same thing, [p] and [b] cannot be phonemically equivalent. However, /lad/ means the same as /lad/ regardless of whether [l] is produced with the tip of the tongue touching the upper dental ridge or with it displaced laterally (as for the second /l/ of the word *laterally*). With such a definition, a phoneme is an emergent property of a lexicon.

At about one year, infants make most phonetic discriminations, and yet, as far as is known, their lexicon is virtually empty. How do they map their phonetic knowledge onto a phonemic system? Why not store /spit/ and /phit/ as different words with different meanings? The answer is that the

child does not wait to compile the lexicon before deciding whether a phonetic contrast is or is not useful in the language.

The evolution of the categorization of speech sounds by infants has been well documented. Indeed, the superb universal aptitudes for categorization displayed by infants at birth tend to decay after some 8 to 10 months of life according to Werker and Tees (1984). Before they turn 12 months of age, infants begin to neglect contrasts that are not used in their linguistic environment. Thus 10-month-old infants raised by English speakers start to neglect the aspirated/nonaspirated contrast used in Hindi. Likewise, other contrasts not relevant to their linguistic environment are lost at roughly the same age. This loss seems to be due to cognitive and attentional processes and not to the loss of sensory-neural sensitivity. If the native and foreign contrasts are similar, unfamiliar speech sounds are assimilated to familiar native categories, which causes the foreign distinction to disappear. In contrast, when unfamiliar speech sounds are too different to be assimilated to familiar native speech categories (Zulu clicks), they are still perceived categorically by inexperienced adults (Best, McRoberts, and Nomathemba 1988). Notice, however, that in all these studies, unfamiliar sounds are phonemically perceived. This suggests that the child has mastered the phonological regularities of his or her language before compiling a lexicon.

If Werker and Tees are right, by the time infants are about to speak, they have already distinguished the relevant from the irrelevant contrasts in their language. Thus the convergence is likely due to a specialized modular system. It seems plausible that computation of the distributions is sufficient to extract statistical modes in speech sounds from the environment.[5] However, no model has as of yet really succeeded in simulating the child's performance. Whatever the exact mechanism of the phonemic shift may be, this reorganization should play a role within future models.

Syllabic and prosodic convergence

Cutler et al. (1986) claimed that English and French adults use different segmentation routines. Indeed, French subjects are sensitive to the syllabic boundary of words like *ba-lance* and *bal-con*, whereas English speakers are not (see Segui, Dupoux, and Mehler, this volume). How does the infant converge on one or the other of these strategies? Empirical investigation of this question with young infants is not yet available. However, Cutler et al. have addressed this issue by testing adult English-French bilinguals. The subjects were raised from birth in a perfectly bilingual environment and were rated as native speakers in both languages. Subjects rated them-

selves as English or French dominant. French-dominant subjects showed a clear syllabification strategy when listening to French materials but not when listening to English materials. In contrast, English-dominant subjects did not show syllabification in either language. The authors interpret this result by saying that there are syllabic and nonsyllabic strategies for parsing the speech signal.

The question that arises is how the organism tunes in the cues that help segment the speech signal into words. Here again, phonotactic and metrical properties of a language can be seen as a by-product or a *conspiration* (McClelland and Elman 1986) of the lexical items stored in the adult's lexicon. Yet how does the child ever manage to store a few words if he does not know what type of segmentation strategy is useful for his language?

Dresher and Kaye (1990) explore how part of the problem might be solved. These authors have shown that given the syllabic structures of individual words and their stress patterns, a computational model can learn the metrical rules of the language. Thus, given a handful of words sufficiently representative of the language, the algorithm decides whether the language is stress initial, terminal stress, etc. The algorithm selects a system from 200 possible metrical systems by scanning the input words for some robust cues that allows it to set the correct combinations of 10 metrical parameters.

This result is all the more interesting in that it meshes well with what is known about the processing abilities of the child: it seems that children do have access to a syllabic representation and that they do take stress patterns into account (Gleitman and Wanner 1982). Moreover, since the database for inferring the metrical rules is not huge, it is quite reasonable to assume that children rapidly attain a metrical grammar for their language from isolated words or with the help of a rough word-boundary detector. Once compiled, a metrical grammar may provide an important cue signaling language-specific word boundaries to the child.

A Hypothesis about Acquisition and Accessing the Lexicon

What is the link between word recognition and the acquisition of a lexicon? In this section we will propose a framework that we call SARAH (syllable acquisition, representation, and access hypothesis). SARAH provides a common theoretical vocabulary for the study of adult and infant speech perception. This framework postulates a strong correspondence between the processes used by the young infant and those underlying lexical access in the adult.

Our framework postulates a structural, coarse-grained, and weakly linguistic unit of processing. It identifies a syllablelike prelexical segment used to construct potential lexical entries at the initial state and to mediate lexical access and phonemic extraction at the stable state. In our model we propose a set of constraints on how lexical *entries* for spoken words are elaborated during the first years of life. SARAH is a model of how the *forms* of spoken words are acquired. SARAH is *not* a model of how the *content* of the mental lexicon is acquired and represented.

The stable state

SARAH claims that in adults, the front end of speech perception relies on three levels of processing: syllabic, lexical, and phonological.

Syllabic level SARAH posits that speech is segmented into elementary units that roughly correspond to the syllable. A *syllabic frame* corresponds to an elementary speech utterance, that is, to the minimal functional unit relevant to speech production. This unit captures invariance across speaking rates. Since the number of syllabic frames relevant for a language is rather reduced (about 6,000 for French), it is reasonable to propose that syllabic frames are recognized by a bank of syllabic analyzers.

Lexical level The bank of syllabic analyzers becomes the code used to access the lexicon. The first syllable of an item constitutes the access code, i.e., the minimal amount of information that can activate a cohort of word candidates.

Phonological level SARAH claims that phonemes do not play a direct role in speech perception but are derived from the prelexical code, namely, from the syllabic frames.

In this volume Segui et al. argue that such a model accounts for many results in psycholinguistic research. However, more research is needed to specify the exact nature of syllabic frames. Whether the unit corresponds to the syllable as defined in phonological theory or to diphones, triphones, wickelphones, or whatever is a question open to empirical investigation, and the answer may be language-specific.

The initial state

SARAH claims that infants are equipped with at least three basic routines that make acquisition of a lexicon possible.

Syllabic filter SARAH postulates a syllabic filter that chops continuous speech into isolated elementary syllabic segments. This filter allows only

legal sequences such as CV, CVC, V, and CCVC syllables to be analyzed. Sounds like [pst] or a dog's bark are never included as basic sounds. These syllablelike segments are specified in an abstract format, where the type of speaker and speech rate have been factored out. These syllables can be looked on as a kind of gestalt, like, for instance, a square or a triangle.

Phonetic analyzer SARAH proposes that the infant uses each syllabic representation to compute the underlying universal phonetic representation of the segment. This phonetic analyzer provides a description of the whole syllable in a format compatible with production routines. We thus postulate that these syllables are mapped onto a code that captures the gestures necessary to produce the syllable. To compare this to visual prototypes, the analyzer extracts elementary descriptions of the figure (curved/straight segments, vertical/horizontal main axis, junctures, etc.) relevant to a production routine (a tracer).

Word-boundary detector SARAH speculates that a word boundary detector uses syllabic representations and other acoustic information to compute elementary cues (duration, stress, etc.) that indicate the onset and offset of words.

This model of the initial state is compatible with findings that suggest that a baby represents speech both at the syllabic and the phonetic levels. Moreover, it acknowledges that infants are able to use prosodic properties of utterances and extract sentence and clause boundaries. SARAH predicts that syllable and word boundaries may also be represented in the signal.

The transition

SARAH claims that the transition from a phonetic to a phonemic system uses two types of specialized mechanisms:

Unlearning, or selective stabilization At the initial state the young infants possess a universal device to process speech. For instance, their phonetic aptitude is such that they can discriminate any linguistic contrast found in any existing language. Specialization is a process by which such a *universal phonetic capacity* is restricted to a set of language-specific contrasts. SARAH predicts that the same is true for the other devices, namely the syllabic filter and the word-boundary detector. Presumably the syllabic level is much richer in the initial state than in the stable state. By specialization, the syllabic level retains only syllables permissible in the native language(s). Finally, an infant retains only the *word segmentation strategy* that works best for his or her language. Such

a process of specialization should depend not on the prior acquisition of a lexicon but rather on such specialized modular routines as statistical extraction and parameter setting.

Compilation Another acquisition mechanism that SARAH proposes is the storage of syllabic templates and logogens into long-term memory. If the other devices used by the infant (the syllabic filter and the word-boundary detector) are functional and properly tuned, this process is accomplished quite automatically without reference to higher-level constraints. The set of syllable templates helps to bootstrap the acquisition of lexical entries. This set of potential words in turn bootstraps higher levels that facilitate the establishment of a proper morphology for the language in question.

We insist that during the early stages of speech acquisition the entries constructed by the child may be quite remote from what they are in adults. For instance, it seems very likely that children acquire their first potential words as syllables and then as clusters of syllables taken from sentences. These may include, in addition to words, some extra materials such as adjacent closed-class words or clitics ([ðedog], or [delo] (*de l'eau*) in French). How these entries are cleaned up to converge toward those of adults, although not yet understood, at least requires the joint operation of bottom-up and lexical-morphological indices.

Concluding remarks about SARAH

Most models of speech recognition have focused on the performance of adults. Yet researchers have reached no general consensus as to the specific units used to represent and process speech. We claim that much is to be gained by changing perspective so as to acknowledge strong constraints that can be derived by taking acquisition into account.

The basic structures and routines that operate in adult speakers must come from somewhere. It is unlikely that all structures come from thin air or from the linguistic environment of the infant. In fact, as many linguists, psychologists, and others have pointed out (Chomsky 1957, Fodor 1983, etc.), extreme instructivist models of language acquisition are inadequate. Speech acquisition is fast, it honors biological landmarks, and its pace is relatively immune to environmental variations. In a sense, speech acquisition is like specialized acquisitions in other species due mostly to instinctive learning, parameter setting, or hypothesis testing. Thus the main structures used by the adult depend on the biological makeup of the species and should be linked to information-processing

structures present in the infants. The rough sketch we have presented has many loose ends and shortcomings. However, it raises many empirical issues and opens new avenues for modeling. We hope that with further research SARAH will acquire speech better than the famous ape of the same name.

Acknowledgments

This research was carried out with the help of CNET (convention no. 00790 9245 DIT), CNRS (ATP Aspects Cognitifs et Neurobiologiques du Langage), and the European Science Foundation (TW 86/17).

Notes

1. We acknowledge that the senses transduce information continuously. The basilar membrane, for one, is set in motion by the acoustic waves in the environment, and information travels up the acoustic fibers in a complex but basically continuous fashion. This mechanism is elegantly illustrated in Delgutte 1980 and Delgutte and Kiang 1984. Likewise, the retina transmits information continuously when the viewer is stimulated by light energy. So there needs to be little, if any, discussion about the continuous versus discontinuous nature of information *transduction*. Our aim, however, is to understand how continuous sensory information is used in speech perception and how the mapping between acoustic energy and higher levels of representation takes place.

2. However, since phonemes are defined over a configuration of distinctive features and since the information flow between different nodes is rather smooth, it could be argued that TRACE *behaves* like a continuous fine-grained model.

3. Of course, a more indulgent reading of this proposition is that the exact duration of the processing window is not fixed but depends on the speech rate. However, such a proposition should specify in a principled way how the speech rate affects the window. In our view this amounts to positing a structural unit.

4. Since the percolation of information into units is related to the rate of variability in the signal, even a structural version of a very fine-grained model is inadequate. First of all, no information is provided as to the value of the individual loops. Thus very odd noises could be learned as proper words (but a burst spectrum of 10,000 msec followed by a voicing onset spectrum of 20,000 msec before a single vowel midpoint spectrum is not likely to be identified as /ta/ by humans). Second, the duration of spectral components is crucial for consonant identification: voice onset time (VOT), the slope of formant transitions, etc. Self loops at every critical band spectrum are thus likely to disrupt performance. As a result, both potential solutions have shortcomings. However, part of the problem might be overcome by using larger units.

5. Data from Lisker and Abramson 1970 indicate that the statistical distribution of VOT in production of the [ba]–[pa] continuum is affected by the native speaker's

language. Clear modes of production are observed in the region of phonemic categories.

References

Abramson, A. S., and Lisker, L. 1965. Voice onset time in stop consonants: Acoustic analysis and synthesis. In *Proceedings of the Fifth International Congress of Acoustics*, Liège.

Bahal, L., Bakis, R., Cohen, P., Cole, A., Jelinek, F., Lewis, L., and Mercer, R. 1981. Speech recognition of a natural text read as isolated words. In *Proceedings of the ICASSP*, Atlanta, Ga., March–April, pp. 1168–1171.

Bahrick, L. E., and Pickens, J. N. 1988. Classification of bimodal English and Spanish language passages by infants. *Infant Behavior and Development* 11:277–296.

Bertoncini, J., Bijeljac-Babic, R., Blumstein, S., and Mehler, J. 1987. Discrimination in neonates of very short CV's. *Journal of the Acoustical Society of America* 82:31–37.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L., and Mehler, J. 1988. An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General* 117:21–33.

Bertoncini, J., and Mehler, J. 1981. Syllables as units in infant speech perception. *Infant Behavior and Development* 4:247–260.

Bertoncinci, J., Morais, J., Bijejac-Babic, R., McAdams, S., Peretz, I., and Mehler, J. 1989. Dichotic perception and laterality in neonates. *Brain and Language* 37: 591–605.

Best, C. T., McRoberts, G. W., and Nomathemba, M. S. 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, no. 3:345–360.

Bijeljac-Babic, R., Bertoncini, J., and Mehler, J. In preparation. Discrimination de séquences multisyllabiques naturelles chez le nouveau-né de quatre jours.

Chodorow, M. S. 1979. Time compressed speech and the study of lexical and syntactic processing. In W. E. Cooper and E. T. C. Walker (eds.), *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*. Hillsdale: L. Erlbaum.

Chomsky, N. 1957. *Syntactic Structure*. The Hague: Mouton.

Church, K. W. 1987. Phonological parsing and lexical retrieval. *Cognition* 25:53–70.

Colombo, J., and Bundy, R. S. 1983. Infant response to auditory familiarity and novelty. *Infant Behavior and Development* 6:305–311.

Cutler, A., Mehler, J., Norris, D., and Segui, J. 1983. A language-specific comprehension strategy. *Nature* 304:159–160.

Cutler, A., Mehler, J., Norris, D., and 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., and Norris, D. 1979. Monitoring sentence comprehension. In W. E. Cooper and E. C. Walker (eds.), *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*. Hillsdale: L. Erlbaum.

DeCasper, A. J., and Fifer, W. P. 1980. Of human bonding: Newborns prefer their mother's voices. *Science* 208:1174–1176.

Delgutte, B. 1980. Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. *Journal of the Acoustical Society of America* 68:843–857.

Delgutte, B., and Kiang, N. Y. S. 1984. Speech coding in the auditory nerve. Parts 1–5. *Journal of the Acoustical Society of America* 75:866–918.

Dresher E. and Kaye, J. 1990. A computation learning model for metrical phonology. *Cognition* 34, no. 2.

Dupoux, E., and Mehler, J. 1990. Monitoring the lexicon with normal and compressed speech: Frequency effects and the prelexical code. *Journal of Memory and Language* 29, no. 3: 316–335.

Eimas, P. D. 1974. Auditory and linguistic processing of cues for place of articulation by infants. *Perception and Psychophysics* 16, no. 3: 513–521.

Eimas, P. D. 1975. Auditory and phonetic coding of the cues for speech: Discrimination of the (r-l) distinction by young infants. *Perception and Psychophysics* 18, no. 5: 341–347.

Eimas, P. D., Siqueland, E. R., Jusczyk, P. W., and Vigorito, J. 1971. Speech perception in infants. *Science* 171:303–306.

Fernald A., and Kuhl, P. 1987. Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development* 10:279–293.

Fodor, J. A. 1983. *The Modularity of Mind*. Cambridge: MIT Press.

Gleitman, L. R., and Wanner, E. 1982. Language acquisition: The state of the state of the art. In E. Wanner and L. R. Gleitman (eds.), *Language Acquisition: State of the Art*. New York: Cambridge University Press.

Halle, M., and Vergnaud, J. R. 1987. *An Essay on Stress*. Cambridge: MIT Press.

Hirsh-Pasek, K., Kemler Nelson, D. G., Jusczyk, P. W., Cassidy, K. W., Druss, B., and Kennedy, L. 1987. Clauses are perceptual units for young infants. *Cognition* 26:269–286.

Jusczyk, P. W., and Derrah C. 1987. Representation of speech sounds by young infants. *Developmental Psychology* 23:648–654.

Jusczyk, P. W., Hirsh-Pasek, K., Kemler Nelson, D. G., Kennedy, L. J., Woodward, A., and Piwoz, J. Submitted. Perception of acoustic correlates to major phrasal units by young infants.

King, P. E., and Behnke, R. R. 1989. The effect of time-compressed speech on comprehensive, interpretive, and short-term listening. *Human Communication Research* 15, no. 3: 428–443.

Klatt, D. 1977. Review of the ARPA Speech Understanding Project. *Journal of the Acoustical Society of America* 62, no. 6.

Klatt, D. 1989. Review of selected models in speech perception. In W. D. Marslen-Wilson (ed.), *Lexical Representation and Process*. Cambridge: MIT Press.

Lachter, J., and Bever, T. G. 1988. The relation between linguistic structure and associative theories of language learning—A constructive critique of some connectionist learning models. *Cognition* 28:195–247.

Lenneberg, E. 1967. *Biological Foundations of Language*. New York: Wiley.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. 1967. Perception of the speech code. *Psychological Review* 74:431–461.

Liberman, A. M., and Mattingly, I. G. 1985. The motor theory of speech revised. *Cognition* 21:1–36.

Lisker, L., and Abramson, A. S. 1970. Some experiments in comparative phonetics. In *Proceedings of the Sixth International Congress of Phonetic Sciences*. Prague: Academia.

McClelland, J. L., and Elman, J. L. 1986. The TRACE model of speech perception. *Cognitive Psychology* 18:1–86.

Marslen-Wilson, W. D. 1984. Function and process in spoken word recognition. In H. Bouma and D. G. Bouwhuis (eds.), *Attention and Performance vol. 10, Control of Language Process*. Hillsdale: L. Erlbaum.

Marslen-Wilson, W. D. 1987. Functional parallelism in spoken word recognition. *Cognition* 71:71–102.

Marslen-Wilson, W. D., and Tyler, L. K. 1980. The temporal structure of spoken language understanding. *Cognition* 8, no. 1: 1–71.

Mehler, J. 1981. The role of syllables in speech processing: Infant and adult data. *Philosophical Transactions of the Royal Society of London* B295:333–352.

Mehler, J., Bertoncini, J., Barrière, M., and Jassik-Gerschenfeld, D. 1978. Infant recognition of mother's voice. *Perception* 7:491–497.

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

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., and Amiel-Tison, C. 1988. A precursor of language acquisition in young infants. *Cognition* 29:143–178.

Mehler, J., Segui, J., and Frauenfelder, U. 1981. The role of the syllable in language acquisition and perception. In Terry Myers, John Laver, and John Anderson (eds.), *The Cognitive Representation of Speech*. Amsterdam: North-Holland Publishers.

Miller, J. L., and Eimas, P. D. 1983. Studies on the categorization of speech by infants. *Cognition* 13:135–165.

Miller, J. L., and Liberman, A. M. 1979. Some effects of later occuring information on the perception of stop consonants and semivowels. *Perception and Psychophysics* 25:457–465.

Mills, M., and Meluish, E. 1974. Recognition of the mother's voice in early infancy. *Nature* 252:123–124.

Piaget, J. 1977. *La construction du réel chez l'enfant.* 6th ed. Paris: Delachaux et Niestlé.

Pinker, S., and Prince, A. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28:73–193.

Port, R. F. 1976. The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochee words. Doctoral dissertation.University of Connecticut.

Port, R. F., and Dalby, J. 1982. Consonant/vowel ratio as a cue for voicing in English. *Perception and Psychophysics* 32:141–152.

Salasoo, A., and Pisoni, P. A. 1985. Interaction of knowledge sources in spoken word recognition. *Journal of Memory and Language* 24:210–231.

Segalowitz, S. J., and Chapman, J. S. 1980. Cerebral asymmetry for speech in neonates: A behavioral measure. *Brain and Language* 9:281–288.

Summerfield, A. Q. 1975. Aerodynamics versus mechanics in the control of voicing onset in consonant-vowel syllables. In *Speech Perception*, no. 4, Department of Psychology, Queen's University of Belfast.

Trehub, S. E. 1976. The discimination of foreign speech contrasts by infants and adults. *Child Development* 47:466–472.

Treiman, R. 1983. The Structure of spoken syllables: Evidence from novel word games. *Cognition* 15:49–74.

Tyler, L. K. 1984. The structure of the initial cohort: Evidence from gating. *Perception and Psychophysics* 36:217–222.

Tyler, L. K., and Wessels, J. 1983. Quantifying contextual contributions to word-recognition processes. *Perception and Psychophysics* 34:409–420.

Werker, J. F., and 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.