CHAPTER

27 Spoken Word Recognition

James S. Magnuson, Daniel Mirman, *and* Emily Myers

Abstract

Spoken word recognition is the study of how lexical representations are accessed from phonological patterns in the speech signal. That is, we conventionally make two simplifying assumptions: Because many fundamental problems in speech perception remain unsolved, we provisionally assume the input is a string of phonemes that are the output of speech perception processes, and that the output is a string of recognized words that are passed onto mechanisms supporting, for example, sentence processing. These kinds of assumptions allow psycholinguists to break language understanding into manageable research domains, and as we review, they have afforded great progress in understanding spoken word recognition. However, we also review a growing body of results that are incompatible with these assumptions: Spoken word recognition is constrained by subphonemic details and top-down influences from higher level processing. We argue that these findings are incompatible with current theoretical frameworks, and that a new theoretical paradigm is needed.

Key Words: lexical access, computational models, segmentation, embedded word problem, interaction, top-down influences

When we comprehend spoken language, we accomplish many amazing feats. We map the acoustics of speech onto phonological categories, despite a rapid signal,[1] with tremendous variation in talker characteristics, acoustic environment, speaking rate, and many other dimensions (see Chapter 26, this volume). We also build syntactic structures on the fly as we hear speech, allowing us to derive a talker's intended meaning, overcoming rampant syntactic and semantic ambiguity along the way (Trueswell & Tanenhaus, 1994). The study of word recognition focuses on processes that intervene, more or less, between phonological and syntactic processing: how do listeners map phonological forms onto words in memory, taking into account prior probabilities, linguistic and nonlinguistic context, including the phonetic, phonological, lexical, semantic, and syntactic structures that have been activated by the speech signal as it unfolds in time. As we shall see, there may not be distinct processes or representations that are concerned purely with spoken word recognition.[2] Word recognition is influenced by (and in turn influences processing of) low-level acoustic and phonetic details of speech as well as semantic and syntactic processing; there is no sharp boundary between speech perception and word recognition, nor between word recognition and sentence processing.

Nonetheless, psycholinguistics is divided into intuitively identifiable levels of organization in human language processing—*speech perception, spoken word recognition, sentence processing*, and so on—providing a logical division of labor among psycholinguists. This way, rather than waiting until all fundamental problems at the level of speech perception are solved, researchers can make progress on the word level by making simplifying assumptions about the nature of the input and the goals of word-level processing (see Magnuson, 2008, for a detailed discussion). The nature of the input and the goals of processing are what David Marr (1982) called the *computational level* of theorizing about information processing systems. We need to articulate a computational-level theory before we can discuss what sorts of representations and mechanisms support human spoken word recognition.

The Computational Problem

Marr (1982) described three levels of information processing theories: computational, algorithmic, and implementational. The idea is that when we are faced with some sort of information processing system—a thermostat, a calculator, a computer, or human language processing—and want to figure out through empirical investigation what the system does and how it works, we need to posit theories at multiple levels of analysis to guide our investigations. The most abstract level is the *computational* theory, which is an analysis of the goals or purpose of the system based on the mapping between the system's input and output—which defines the information processing the system does. For a basic calculator, the input might be numbers and symbols for mathematical operations, and the output would be the result the calculator displays.

At the *algorithmic* level, we develop a theory of the "software" that could achieve the computational mapping. For example, we might posit that for multiplication, the calculator performs a sequence of additions (given 4 x 3, it adds 4 + 4 + 4) or that it just looks up the result from a list of pairwise products coded in its memory. We might then propose an empirical test of these two theories by measuring how long the calculator takes to display its answer in cases where the two theories would require a similar number of operations versus cases where the serial addition theory would involve many more operations.

The third level is the *implementational* level, which in the case of our calculator would be the hardware level. Here, we examine the inner workings of the calculator—perhaps taking it apart, enumerating the components and their interconnections. In a modern calculator, this would be a circuit board. We take as given that the relevant implementational level for spoken word recognition is that of neural systems. Despite the advent in recent years of sophisticated neuroimaging techniques, understanding of the neural basis of language remains rather coarse (for recent reviews, see Blumstein & Myers, in press; Hickok & Poeppel, 2007; Ueno, Saito, Rogers, & Lambon Ralph, 2011). While we argue at the end of the chapter that integration with cognitive neuroscience via lesion studies, neuroimaging, and genetics will be necessary for full understanding of human language processing, our focus will remain firmly on cognitive psychology in this review.

One of Marr's goals in identifying these three levels of information processing theory was to make clear that they can be addressed independently. Theories at different levels need not have contact with each other to make progress (although a computational level theory—the relevant input-output mapping and its purpose—is a prerequisite to an algorithmic theory and a crucial source of constraints in studying the implementational level), but they can be mutually constraining. For example, if we are comparing algorithmic theories of multiplication that posit serial addition vs. memory lookup (for single digits), we might search for circuits at the implementational level that perform addition when a multiplication operation is called for. Conversely, if we determine at the implementational level that the device has a memory limited to 1 kilobyte, we should not propose algorithmic theories that require megabytes of memory.

Now imagine we have a device that happens to be a calculator—but we do not know that, because the Arabic numerals and mathematical symbols have been replaced with arbitrary symbols. What should our first step be in figuring out what information processing this device does? We will have to observe what symbols are displayed as we input different key sequences. We should arrive at the same computational theory as we would with a "normal" calculator, but it will take a lot of work to document the input-output mappings and surmise the purpose of the information processing the device does. Enumerating all possible inputs and their results would not be feasible—depending on the limits of the calculator, these may be infinite or at least impractically many. If we have no clue that this device is a calculator, imagine how difficult it would be to jump to the algorithmic or implementational level before establishing a computational theory—that is, figuring out that the purpose of the device is calculation. And this is largely the position we find ourselves in when attempting to develop theories of cognition. So let's begin by deriving a computational theory of *spoken word recognition*: What are the inputs, the outputs, and the mapping between them?

The input is a trickier issue than one might think. It may seem obvious that for *spoken* word

recognition, the input is speech. This turns out to be an impractical starting point, because of unsolved mysteries in speech perception. The central one is called the *lack of invariance problem*, and it refers to the fact that there is not a simple mapping between acoustic patterns and percepts (such as consonants and vowels, syllables, or words). Instead, there is a many-to-many mapping (e.g., Lisker, 1985). The acoustic patterns that map to a particular phoneme (consonant or vowel) vary depending on phonetic context (what phonemes precede and follow), acoustic environment (an open field vs. an echoey stairwell vs. a noisy party), talker characteristics (physical size, sex, dialect, idiolect, emotional state), speaking rate ... and this is a partial list! (See Chapter 26, this volume, for a review.) The result is that many acoustic patterns can map to the same percept. Compounding this problem is the fact that phonological categories overlap in acoustic space and an identical acoustic pattern can map onto multiple percepts depending on context. For instance, a given vowel can sound like the vowel in "bit" in the context of one talker, but like the vowel in "bet" in another talker's voice (Ladefoged & Broadbent, 1957; Peterson & Barney, 1952), and the same acoustic pattern can sound like "b" in the context of slow speech but "w" in fast speech (Miller & Liberman, 1979; for examples of how speech rate affects spoken word recognition, see Dilley & Pitt, 2010).

The acoustic variability in different phonetic contexts arises due to the fact that speech is *coarticulated*, meaning that the vocal tract gestures we use to produce successive speech sounds overlap. If we try to identify the time intervals where the vocal tract is making the motor movements for the phonemes /b/, /æ/, and /g/ in the word "bag," we find that the gestures for /æ/ extend over the entire production of the word, and even gestures for /b/ and /g/ overlap in time (for a review, see Fowler & Magnuson, 2012). Coarticulation contributes to the lack of invariance problem (the acoustics for a phoneme will change given the motoric constraints of producing adjacent phonemes with which it is coarticulated) but is also at the heart of the *segmentation* problem: There are few acoustic cues to where one phoneme ends and another begins because articulation and therefore acoustics are usually continuous. Discontinuities—actual breaks and pauses in the signal, which would seem like a logical acoustic cue to boundaries—are not reliable cues to phoneme boundaries. These discontinuities rarely occur "between" phonemes and are likely to happen "within" a phoneme, such as the sudden reduction in sound when a constriction for a voiceless stop consonant like /p/ is made (the constriction gesture being the brief, air-stopping closure of the lips).

Rather than waiting until these mysteries are resolved, psycholinguists make simplifying assumptions that allow them to tackle higher levels of processing. Typically, psycholinguists allow themselves to assume that the input to word recognition will be a string of phonemes generated by a speech perception mechanism (and scientists studying sentence processing make a similar assumption and assume a string of words to be a plausible input to sentence processing—and so on, as one moves to higher levels of organization). This simplifying assumption allows us to get started, but as we shall see later in this chapter, it paradoxically makes some aspects of theories of spoken word recognition more complex than they need be.

For now, let's consider what this assumption buys us. If we consider the input to be a string of phonemes, the next step in deriving a computational-level theory of spoken word recognition is to ask what the output should be, in order to identify the *goal* of the information processing system we are studying. The ultimate goal is, of course, a full specification of the listener's understanding of the speaker's message, but the complexities of semantics and syntax render this, as with the input, too large a problem to take on simultaneously with word-level processing. We would also have to grapple with how words map onto syntactic and semantic structures of various sorts, how those result in appreciation of the speaker's intent (or not) in light of linguistic and nonlinguistic context, the listener's prior experiences, and so on. Rather than grappling with all of these details in one go, psycholinguists break the problem into manageable chunks. For spoken word recognition, we assume the intermediate goal of the system is to achieve a match to a *lexical representation* of a word in memory and feed it forward for higher levels of processing.

What should be the contents of lexical representations? Certainly, there will have to be a *phonological form*—a sequenced list of the phonemes that should occur (the most typical assumption) or some other acoustic form (e.g., episodic traces; Goldinger, 1998), or detailed phonetic or phonological form. While one possibility we have just listed is explicitly nonphonological (acoustic episodes), we will use "phonological form" to mean any coding scheme that maps the speech signal onto words in the lexicon. While phonological form will be the primary key for accessing lexical representations, what else should we assume lexical representations contain? We might minimally expect *lexical semantics* and *grammatical class*, although contemporary linguistic and psycholinguistic theories of grammar and sentence processing attribute even more syntactic and semantic knowledge to lexical representations (Altmann & Mirković, 2009; MacDonald, Pearlmutter, & Seidenberg, 1994; Pusteyevsky, 1995; Trueswell & Tanenhaus, 1994). Again, theories of spoken word recognition typically defer consideration of details beyond phonological form to higher levels, such as sentence processing. As with input representation, the risk of seemingly simplifying assumptions may make the problem of spoken word recognition paradoxically more difficult if they exclude contributions of semantic, syntactic, and pragmatic factors to resolving word-level ambiguities (e.g., Barr, 2008; Dahan & Tanenhaus, 2004; Magnuson, Tanenhaus, & Aslin, 2008).

So now we have a basic definition of the computational problem: The input is a string of speech sounds represented either abstractly (as symbolic representations) or episodically (as memory traces preserving surface details),[3] and the goal is access of a lexical representation, which includes at least parsing the phonemes onto phonological forms of words in memory, and may include accessing some degree of semantic, syntactic, and pragmatic knowledge (though these details tend to be beyond the scope of current theories of spoken word recognition). The next steps in determining how the system achieves this goal are characterizing the mappings between inputs and outputs through experimental observation, and then enumerating the computational challenges the system must overcome in achieving the mapping, and constraints on how it does so (How quickly are particular words processed? What kinds of errors do listeners make?). With a sufficiently large body of results of this sort in hand, one can begin to construct algorithmic ("software") level theories specifying cognitive mechanisms that could achieve the observed mapping. This sets the stage for competition among algorithmic theories, as theories generate experimentally testable predictions that go beyond the phenomena that have been observed so far, and the knowledge base and level of detail about the input-output mapping grows. Thus, as algorithmic theories are posited and tested, the computational theory—the mapping that any algorithmic theory must account for—becomes more detailed and refined.

We will use an even more abstract lens—Kuhn's (1962) three phases of scientific paradigms—to present our view of the literature and history of spoken word recognition: *preparadigm*, where basic facts are enumerated and preliminary, incomplete theories begin to point the way to a theoretical consensus; *normal science*, where theories drift toward consensus in response to an ever-growing corpus of empirical details, but eventually anomalous findings that cannot be accommodated by extant theories push the paradigm toward a crisis; and *revolutionary science*, where a new paradigm arises from insights that allow for a new theoretical perspective that accommodates the anomalies (a classic example from Kuhn is Gallileo's insight that friction can explain why terrestrial objects in motion stop without some force pushing them, rather than this being an intrinsic natural property of objects, which paved the way for a paradigm shift to Copernican cosmology). With this framework in mind, we will next present a selective and not purely chronological review of the spoken word recognition literature in three parts: the foundational, preparadigm consensus (the initial computational-level facts on which spoken word recognition researchers agree); the progression to an initial phase of normal science, where models and measures of the time course of processing pushed theory development; and finally, where we find ourselves today—(hopefully) late in a period of normal science, with a growing body of anomalous findings that cannot be easily fit into current theories. We will conclude the chapter with a discussion of what we view as the most pressing crises, and possible avenues to a revolution. These include grappling with the actual speech signal, integration of word-level processing with the context of language understanding, development throughout the life span, and neurobiological constraints on algorithmic theories.

Basic Science Phase: Foundations of a Computational Theory of Spoken Word Recognition

To set the stage for our review of foundational empirical results, the basic facts on which all theories of spoken word recognition agree are summarized in Table 27.1.[4] These basic facts were discovered in the middle- to late-1900s as the essential technologies, methods, and theoretical framework emerged. These developments, namely the advent of chronometric (i.e., precise reaction time) experimental methods, digital speech analysis and manipulation methods, and the development of theories of

Table 27.1 Basic, Agreed-Upon Facts About Spoken Word Recognition

Spoken Word Recognition Is...	Details, Implications, Challenges, Etc.
Incremental	As a word is being heard, that is, as soon as even the initial sound is heard: ➢ Multiple words are activated in *parallel* in memory, ➢ With strength proportional to their *similarity* (both phonetic and semantic) to the input and *prior probability* (*frequency* of occurrence and, to a lesser degree, fit to lexical, sentential, or other *contextual constraints*); and ➢ Activated words *compete* for recognition
Sequential	➢ *Coarticulation*: sound patterns corresponding to phonological categories such as consonants and vowels are constellations of temporally overlapping (but not necessarily coincident) buzzes, chirps, and frication that must be bound together to map onto phonological categories ➢ *Phoneme segmentation problem*: phonemes overlap in time and there are no invariant boundary cues[6] ➢ *Lexical segmentation problem*: there are no invariant cues to word boundaries ➢ *Embedding problem*: segmentation must not lead to "recognition" of embedded words (e.g., when hearing *window*, the system should recognize just that word, and not *win*, *wind*, *in*, or *dough*). This is potentially a very large problem; McQueen et al. (1995) estimate that 84% of English polysyllabic words contain at least one embedded word.
Interactive	Spoken word recognition *influences* performance on speech perception tasks and is *influenced by* semantic, syntactic, and pragmatic context.

cognitive psychology, still form the core of spoken word recognition research.[5] Next, we will unpack this consensus, beginning by focusing on phonological form, introducing some of the "basic," "incomplete" theories of the preparadigm phase, and eventually turning to meaning and context.

First, we can intuit the necessity of spoken word recognition being sequential from the *segmentation problem*: Despite our subjective impression that words "pop out" from continuous speech, there are in fact no perfectly reliable cues to word boundaries (Aslin, Woodward, LaMendola, & Bever, 1996; Cole & Jakimik, 1980; Lehiste, 1970); indeed, as we have already reviewed briefly, there are no such things as phoneme boundaries (see Chapter 26, this volume). (To refute your subjective impression that there are breaks between words when listening to a language you know, try discerning word boundaries when listening to someone speak an unfamiliar language.) The absence of robust acoustic boundary cues makes it impossible, for example, for spoken word recognition to depend on a process that buffers the acoustic input until a boundary is detected, and then performs recognition on the entire word form in parallel.

Sequential processing can also be inferred from behavior. The first time-course method devised to study spoken word recognition was the *gating task* (Grosjean, 1980). In gating, a small portion of the onset of a word—the first "gate"—is presented, and the subject is asked for the most likely completion. Then, the second gate—a slightly longer portion of the word, beginning from word onset—is presented, and the subject guesses at a completion. This continues with successively longer gates until the entire word is presented. Recognition is operationalized as the gate by which the correct word is always given. Even at the first gate, participants are able to offer completions. As more of the word is presented, the number of completions offered decreases, in a fashion consistent with the phonemes heard so far. Given /b/ as the first gate, participants provide completions that begin with /b/. If the second gate provides strong evidence that the second phoneme is /æ/, the completions narrow to words beginning with /bæ/. Gating reveals many additional details about word recognition. For example, word frequency is an important predictor of completion probability (e.g., given /bæ/, *bat* is a more likely response than *bass*). Gating results also suggest that words are often recognized before they have been fully heard; that is, high identification accuracy is often possible before the entire word is presented. This *recognition point* is highly correlated with the *uniqueness point*—the phoneme at which there is only one possible (uninflected) completion of a word, such as the /f/ sound in *elephant*. In some cases, though, the recognition point can even precede the uniqueness point (e.g., when one possible completion is much more probable due to word frequency). Thus, gating suggests that lexical activation is incremental (people are able to provide highly likely completions that are consistent with the gated input), that multiple words are activated (given the variety of completions offered), and that word onsets are crucial keys to accessing lexical items—subjects virtually never provide a completion that mismatches with the initial sounds in the gate (e.g., given /bæ/, subjects do not suggest *cat* as a completion).

So far we have support for (a) sequential/incremental/continuous processing beginning from word onset, (b) multiple activation, and (c) roles for similarity and prior probability (frequency). However, the gating task is rather unusual, and one could argue that it bears little similarity to word recognition "in the wild," and may be instead a guessing game subject to various strategies. Converging evidence, though, comes from tasks like *lexical decision* and *naming*. In a lexical decision task, you hear a spoken word (*ball*) or a spoken nonword (*balt*), and press a button to indicate whether what you heard was a word. In naming (sometimes called *shadowing* or *repetition*), you hear a word and repeat it as quickly as you can. In tasks like these, response latencies decline and/or accuracy increases (especially when the speech signal is degraded or noise is added) with word frequency (e.g., Luce, 1986; Luce & Pisoni, 1998). These measures also provide evidence for activation of and competition among multiple words, as reaction times increase and/or accuracy declines with the number of words in the lexicon that are phonologically similar to a target word.

It is more challenging to detect the sequential nature of processing with tasks like lexical decision and naming. They provide a single, presumably postperceptual measure, and like gating, arguably have little connection to word recognition outside the laboratory. Furthermore, lexical decision has the potential to index something other than the actual recognition of a target word; for example, one might achieve high accuracy just from responding based on a sense of familiarity, or based on the summed activation of multiple words, before one has truly identified the word (Grainger & Jacobs, 1996; Rogers, Lambon Ralph, Hodges, & Patterson, 2004). The coordination of perception and production required by the naming task might direct attention to different details of the signal than a situation where there is no need to repeat a word as you hear it.

However, an extremely clever variant of lexical decision, the *cross-modal semantic priming* paradigm, provides converging evidence for other details from the gating task. The paradigm exploits the phenomenon of semantic priming (Meyer & Schvaneveldt, 1971), where hearing or seeing a related word appears to preactivate or prime semantically related words (e.g., you are faster to recognize *doctor* if it is preceded by *nurse* than if it is preceded by *sandwich*). In cross-modal priming, your task is to perform lexical decision on visually presented letter strings, which are interspersed with auditory stimuli (e.g., Marslen-Wilson & Zwitserlood, 1989). This allows experimenters to look for semantic or other effects between the spoken and written words. In fact, the paradigm might be better called "cross-modal, *phonologically mediated* semantic priming." Rather than looking for direct competition between the spoken prime and visual target, Marslen-Wilson and Zwitserlood (1989) predicted that if a spoken word (e.g., *castle*) activates a cohort of words with similar onsets (*candy*, *cabin*, etc.) to a significant degree, those words should spread detectable activation to semantic relatives (*sweet*, *log*, etc.). If instead word onsets are not crucial, other highly similar words should also be in the competition cohort (and so *castle* should activate *hassle*, which should prime *bother*). The former predictions were borne out: Priming was found for pairs like *castle-sweet*, but not pairs like *castle-bother* (even when word frequency and other factors were controlled). It is important to note that these results also imply that the set of activated words includes items that are semantically related to words that are phonologically consistent with the spoken input. However, the ramifications of this implication are mostly ignored by models of spoken word recognition—a point to which we will return at the end of this section and in the next section.

Marslen-Wilson and colleagues took the gating and priming results to rather transparently reflect the workings of human spoken word recognition and proposed the *Cohort Theory* of spoken word recognition (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978); because their definition of the competition cohort was based on onset overlap, "cohort" has become a synonym for onset competitor.) One of their key insights was that the apparent complication of sequential input actually provides a basis for understanding not just isolated word recognition but also how the system might address the embedding and segmentation problems in isolated words and word sequences.

They proposed a mechanism that at utterance onset begins activating all words consistent with the input. Given the input /b/, all /b/-initial words receive modest activation. As the input continues to /bæ/, all /bæ/-initial words receive more activation, and all /b/-initial words that do not continue to /æ/ are removed from the activation cohort (or begin to be inhibited). This continues as more input comes in. If only a single word is uttered, the latest point at which the word will be recognized is after the final phoneme is heard. However, the word can be very strongly activated prior to word offset as the cohort decreases in size. In the extreme case that a single word remains in the cohort prior to word offset (e.g., /bænIst^r/-*banister*, which becomes unique at /s/), uniqueness point effects are predicted. The original version of the model was interactive. It allowed context to activate words so strongly they could be recognized significantly earlier than their uniqueness points. The model was revised (Marlen-Wilson, 1987, 1989) with strong bottom-up priority—making initial processing autonomous, to avoid the problem that words unexpected in a context can still be clearly recognized (e.g., the system should not recognize *nice* instead of *knife* given *he cut the bread with a nice switchblade*). In the next section, we will discuss how these difficulties with interaction can be avoided when one does not assume there is actually an instant of definitive recognition (the so-called "magical moment"; Balota, 1990), rather than flux in the relative activations of lexical representations.

In the case of a sequence of words, phoneme-by-phoneme winnowing of the cohort provides a potential solution to the segmentation problem (the absence of reliable bottom-up cues to word boundaries). For example, if the utterance so far is /bæn/ (*ban*), and the next sound is /I/ (*ih*, as in *banish* or *banister*), a word boundary cannot be posited. Even though *ban* is a word, longer words beginning with that string remain in the cohort (set of activated words). If instead the next sound is /v/, a word boundary must be posited, because there are no words that begin *banv*. In some cases, a boundary must be posited at an earlier position. For example, if the input has so far been /bænI/ (*bani-*) and the next sounds is /f/ (as in the word sequence *ban if*), a boundary must be posited before the /I/. This algorithm also provides an explanation for how the embedding problem might be handled. In the case of *banister*, *ban* is not recognized because longer words remain in the cohort when /bæn/ has been heard. Thus, segmentation and handling of embeddings emerge from a simple parsing mechanism motivated by the need to handle sequential input and result in a very specific *similarity metric*, that is, the basis for including a word in the activation set. In the case of the Cohort Model, the similarity metric is roughly that two words will compete strongly if they overlap in the first one or two phonemes (Marslen-Wilson, 1987).

But this is where we encounter our first basis for debate about the basic empirical facts for form activation. We cited work by Luce and colleagues earlier in support of the basic fact that recognition facility (speed and accuracy) increases with word frequency and decreases with competitor set size (Goldinger, Luce, & Pisoni, 1989; Luce, 1986; Luce & Pisoni, 1998). While one obtains the same results manipulating the size of the onset cohort (e.g., Zhang, Randall, Stamtakis, Marslen-Wilson, & Tyler, 2011), Luce and colleagues tested a very different similarity metric. This metric, known as neighborhood density, was motivated by the observation that the Cohort Model must assume virtually noise-free input to work as we have just described it (otherwise, the cohort cannot be winnowed accurately), but in the real world, speech is often heard in noisy contexts. So they devised a more forgiving metric that assumes less input certainty and so greater confusability. Extending some previous notions of structure in the auditory lexicon (Greenberg & Jenkins, 1964; Landauer & Streeter, 1973), they proposed what is now called the "DAS" similarity metric: Two words are considered neighbors (and likely to activate each other) if they differ by no more than one phonemic deletion (D), addition (A), or substitution (S). As they predicted, the more neighbors a word has, the more slowly and/or inaccurately it is processed. This is consistent with the notion that neighbors are activated and compete for recognition, and further evidence for competition comes from the result that the neighborhood effects are amplified by neighbor frequency; if two words are matched in frequency and number of neighbors, but the mean frequency of one word's neighbors is higher, that word will be harder to process (Luce & Pisoni, 1998).

This leads to a very different conception of the competitor set compared to Cohort. The word *cat*, for example, has the deletion neighbor *at*, addition neighbors *scat* and *cast*, and substitution neighbors like *bat*, *cot*, and *can*. Many items that would be in the Cohort competitor set are excluded: monosyllables like *camp* and *cask*, and manylonger words (*cabin*, *cabinet*, *cabbie*, *caddy*, *calcium*, *candy*, *catalog*, etc.). Thus, this metric conflicts with the Cohort metric by including words that mismatch at onset, and by excluding many words for which Marslen-Wilson and colleagues reported robust competition effects. Luce and Pisoni gloss over these seemingly problematic items by maintaining their focus on monosyllabic words, acknowledging that aspects of Cohort theory may need to be integrated into their similarity metric later to handle longer words.

Luce and colleagues also showed that the DAS rule can be improved by using a graded similarity metric. Instead of counting complete matches or mismatches, you calculate pairwise similarity between words phoneme-by-phoneme. Luce and colleagues have done this by using actual confusion probabilities measured when speech was presented in noise, but one could also use a metric based on similarity in phonetic features. While this approach provides a modicum of greater precision, Luce and colleagues have reported that the DAS rule works nearly as well, with one important exception: The graded similarity approach predicts inhibitory priming between words that are highly similar at every position but do not overlap in even a single phoneme, such as *veer* and *bull* (these examples differ from each other by a single phonetic feature at each phoneme; see Luce, Golding, Auer, & Vitevitch, 2000, for priming results supporting this prediction).

Luce and colleagues proposed a model based on their results: the *Neighborhood Activation Model* (NAM). NAM does not address the time course of processing—and so is mute on questions of incrementality—focusing instead on multiple, parallel activation and competition. NAM models an assumed final stage of spoken word recognition where acoustic-phonetic detectors have accumulated activation from bottom-up input, which they combine with lexical knowledge, such as word frequencies. Decision unit activations are assumed to be proportional to *frequency-weighted neighborhood probability*; here is our slightly streamlined version of the FWNP:

$$FWNP = \frac{f_t S_{tt}}{\sum f_w S_{wt}} \qquad (1)$$

Thus, the *FWNP* for target word *t* is the ratio of $f_t S_{tt}$ where f_t is *t*'s log frequency, and s_{tt} is *t*'s similarity to an utterance of *t* (which is not necessarily 1.0, as one might confuse the /t/ of *bat* with /d/, for example), to the sum of the similarity to *t* of every word, *w*, in the lexicon (s_{wt}), weighted by its log frequency (f_w). We can simplify further if we base similarity on the DAS rule. Now a word is either a neighbor or it is not, because it either meets the 1-phoneme difference threshold or it does not. Since *s* will be 1 for all neighbors and 0 for every other word, we drop it, leaving the ratio of *t*'s log frequency to the sum of all its neighbors' log frequencies (Equation 2, where *w* has been replaced by *n* in the denominator, as the summation is now over all neighbors, not all words). Because *t* will be included in the denominator, since it is a neighbor of itself, we can think of this as representing the proportion of the frequency weight of its own neighborhood that a word contributes.

$$FWNP_{DAS} = \frac{f_t}{\sum f_n} \qquad (2)$$

This compact, elegant equation (essentially a simplified variant of the R. D. Luce [1959] choice rule) simultaneously represents the core theoretical stance of NAM and provides a crucial methodological tool—studies of spoken word recognition routinely use Equation 2 to control neighborhood (or at least control the raw neighbor counts). Thus, decision unit activations are assumed to be proportional to FWNP, and human behavior is predicted to be proportional to decision unit activations, such that the higher the FWNP for a word, the faster and/or more accurately it should be recognized. The FWNP has largely been tested in the aggregate, with factorial manipulations of FWNP or number of neighbors or with regressions examining how well the FWNP predicts recognition facility for a large number of words (Luce & Pisoni, 1998), as opposed to the typical approach in testing the Cohort model—assessing priming of specific words or enumerating gating completions for specific words. In regression analyses, FWNP accounts for approximately 15% of the variance beyond the 5% accounted for by word frequency alone. We will have more to say about these competing visions of the competitor set in the next section.

Now let's consider evidence for contextual constraints on spoken word recognition. The question is whether semantic, sentential, or other contextual information interacts directly with the bottom-up mapping of sublexical information to lexical representations, or if bottom-up lexical activation is initially *autonomous*—protected from context, which is integrated at a later stage. Frauenfelder and Tyler (1987) made a useful distinction between structural and nonstructural context. Nonstructural context is like word-to-word priming; it does not cross levels of hierarchical organization. If we assume that

semantic and (phonological) form representations exist at the same lexical level, finding priming of the sort we have already described does not address the autonomy issue. Experiments by Tyler, Voice, and Moss (2000) showing that high imageability facilitates recognition of words with many cohort competitors suggests that form and meaning are indeed integral (we will discuss this more in the next section). Evidence of early impact of structural context (e.g., sentence or discourse details), however, would violate bottom-up autonomy. An example of structural context comes from a study by Tanenhaus, Leiman and Seidenberg (1979; see also similar work by Swinney, 1979). They used cross-modal semantic priming, but with auditory stimuli presented in sentence contexts. Their critical items were homophones that were presented in contexts that favored one meaning (*they all rose*) or the other (*he gave her a rose*). Their first question was whether priming would be found for associates of both senses (*stand, flower*). When they presented the probe item visually at homophone offset, they found statistically equivalent priming for associates of both meanings. If they waited 250 msec, there was selective priming for the context-appropriate meaning. Tanenhaus et al. interpreted this as consistent with a mechanism where autonomous, full access to all items matching the bottom-up input is quickly followed by a process integrating the bottom-up signal with sentential context. Shillcock and Bard (1993) used more constraining contexts that strongly predicted closed-class words (e.g., *would*) over open-class homophones (e.g., *wood*) and found selective priming as early as they could look for it (partway through the homophone). This study tends to be neglected in reviews of this literature, but we shall see later that this result has been replicated and extended using newer techniques.

Interaction in the opposite direction—from words to sublexical processing—has been of great interest in spoken word recognition. Examples where lexical knowledge affects sublexical performance include the word superiority effect (Rubin, Turvey, & Van Gelder, 1976), where phonemes can be detected more quickly in the context of a word than nonword, and phoneme restoration (Samuel, 1981, 1996; Warren, 1970), where a phoneme replaced with noise appears to be filled in in a context-appropriate fashion, even having perceptual effects like those of clear phonemes, such as selective adaptation (Samuel, 1997, 2001). While such effects are consistent with the idea that there is direct feedback from words to phonemes, lexical effects on *performance* in phoneme tasks could also arise postperceptually. We will discuss this possibility later; for now, the important thing to note is that theories of speech perception must provide some account of these top-down effects.

The issues we have reviewed have emerged as the primary questions theories of spoken word recognition address—that is, as the boundaries of spoken word recognition theories. In particular, research in the latter half of the 20th century established a framework that assumes that phonemic input activates multiple words in parallel as a function of similarity and prior probability, and activated words compete for recognition. In the next section, we discuss how research over the last couple of decades has filled in many details about these questions, but has also begun to strain at these borders.

The Normal Science Phase: An Emerging Consensus

Here is where this review departs from chronology to discuss the *normal science* phase of contemporary research on spoken word recognition. "Normal science" is what Kuhn (1962) calls the "puzzle-solving" or filling-in period, as (and after) a consensus on a *paradigm* emerges. To mix Kuhn and Marr, this is the consensus on Marr's (1982) computational level theory and agreement on the sorts of experimental methods and measures that provide valid evidence (the most common experimental tasks, along with their advantages and disadvantages are summarized in Table 27.2). On our view, the best way to get a sense of the current paradigm is by walking through the details of the TRACE model of speech perception and spoken word recognition (McClelland & Elman, 1986). This is not to say that TRACE *is* the consensus. However, there is substantial agreement that the *functions* TRACE provides—for example, activation of representations at multiple levels (phonemes, words), inhibition providing the means for competition among activated representations—are needed, even if there is disagreement (and occasional fractious debate) about the best ways to "wire up" those functions (e.g., What is the best similarity metric for predicting what words will be coactive? Is lexical competition better modeled by lateral inhibition between words or bottom-up inhibition from phonemes to words? Should we allow feedback between levels of representation in language processing, or does information flow only in a bottom-up direction?). Most crucially, TRACE ushered in a new level of detail in predictions about not just recognition time but also

Table 27.2 Common Paradigms for Studying Spoken Word Recognition

Task	Advantages	Disadvantages
Lexical decision: Nonwords (e.g., "blat") are mixed with words, and participants press a button indicating YES for words and another indicating NO for nonwords (an alternative "go/no-go" version asks participant to press a button for words and withhold a response for nonwords, or vice versa)	• Fast • Commonly used	• Correct responses do not require full word recognition, but just a relative sense of familiarity or partial activation of multiple words • RT on critical word stimuli is sensitive to design of nonword filler items
Naming (also called Shadowing or Repetition): Participants hear a word and repeat it as quickly as possible (alternative: participants report what word they heard by typing it into a computer keyboard).	• Direct measure of recognition: processing entire word form is required	• Slow: Participants need more time for each trial, so they can complete fewer trials • Does not guarantee deep (e.g., semantic-level) processing • Accuracy requires coding the responses, which can be time consuming and ambiguous (e.g., should mispronunciations/misspellings count as correct or incorrect responses?)
Semantic judgments: Participants indicate whether some semantic property (living thing vs. artifact, something you can touch, etc.) is true of the concept named by the spoken word	• Fast • Requires access to lexical semantic knowledge	• Semantic variables may complicate results (e.g., edible plants such as "tomato" are somewhat ambiguous with regard to their status as living things—when it is on the plant it is living, when it is on the plate it is not)
Word-to-picture matching: Participants indicate which of several pictures matches a spoken word.	• Fast • Requires semantic access • Naturalistic (does not require meta-linguistic judgments) • Can be combined with eye- or hand-tracking to measure the time course of spoken word recognition	• Limited to words that refer to pictureable objects or actions • Reaching movements can make RTs noisy • Sensitive to number of alternatives and their similarities
Priming: Using any of the above tasks, test processing of a word when the preceding word is related on some dimension compared to when it is unrelated. Variants: • Phonological (BALD—BALLS, HAND—SAND) • Phonetic (BULL—VEER) • Semantic (DOCTOR—NURSE) • Cross-modal: prime presented auditorily, target presented visually (or vice versa, depending on which modality the researcher wants to drive initial access to lexical memory)	• Fast • Sensitive • Can measure time course by manipulating the relative timing of the prime and target ("interstimulus interval")	• Priming can be positive (facilitatory—related prime causes better performance) or negative (inhibitory—related prime causes worse performance), and the literature is rife with conflicting positive and negative priming effects • Phonological and semantic priming are susceptible to bias and strategy (e.g., sensitive to proportion of related-prime trials; see Luce et al., 2000, for the relative susceptibility of phonological vs. phonetic priming)

the subphonemic time course of lexical activation and competition. In this section, we will begin with a description of TRACE and the motivation for its architecture and processing components. This sets the stage for debates about the specific mechanisms a model should employ, and the advent of experimental methods that provided time course measures comparable to the time course predictions of models like TRACE. Time course modeling and measures brought about a period of intense research aimed at filling in fine-grained details in the corpus of empirical knowledge (which continues today). In the section following this one, we will argue that current debates are really a matter of fine-tuning the consensus on how one might model the agreed-upon computational theory (that phonemic input activates multiple words in parallel as a function of similarity and prior probability, and activated words compete for recognition)—especially when compared with emerging crises of empirical facts that cannot be accommodated in current theory; a harbinger of a scientific revolution, according to Kuhn (1962).

TRACE and the Time Course of Word Recognition

THE TRACE MODEL OF SPEECH PERCEPTION AND SPOKEN WORD RECOGNITION

Let's walk through how the TRACE model implements the core theoretical consensus (along with some debatable details). TRACE (McClelland & Elman, 1986; McClelland, 1991) was inspired by the Cohort model, but the competition dynamics of the *interactive activation* framework (McClelland & Rumelhart, 1981) were substituted for bottom-up inhibition (inhibiting words that do not include a perceived phoneme) *and* for the notion of an explicit segmentation tracking device. Recall that in Cohort, the system tracks possible completion of words given the input so far; a word boundary is detected when the input so far corresponds to a word and the following segment cannot be added to it. In TRACE, competition and segmentation are emergent properties; there is no explicit tracking of word boundaries. Figure 27.1 shows how this works. On the left, we show a conceptual schematic of the interactive activation framework as it is implemented in TRACE. "Pseudo-spectral" inputs (shown in the center of the figure) activate feature detector nodes, which activate phonemes that contain them. Phonemes send activation forward to words that contain them. Words send feedback to phonemes that they contain. Competition comes from lateral inhibition within the phoneme and word levels. Inhibition among activated units usually leads one node at each level to dominate (achieve the highest activation) for some period of time. Activations wax and wane as a function of bottom-up input, lateral inhibition, and top-down feedback, as the set of strongly activated words changes gradually over time. When a sequence of words is presented, there is nothing inherent in the model that corresponds to a discrete, "magical moment" (Balota, 1990) of word recognition for each word; instead, as the input unfolds, the lexical nodes for the presented words briefly dominate the lexical level in series.[7] To make explicit comparisons with a specific task, such as lexical decision, one must construct a linking hypothesis between the behavior of the model and measures of human performance, such as reaction times (see Magnuson, Mirman, & Harris, 2012).

The actual architecture of TRACE is more complex than the conceptual schematic suggests. A problem that any *implemented* model of spoken word recognition must grapple with is temporal order. Many preceding theories side-stepped this problem; since they were not implemented, they simply stipulated that lexical representations would be sensitive to temporal order (the Cohort Model being a notable exception). Addressing this problem is crucial and very difficult, so it is instructive to walk through how TRACE handles it. If the network were really as simple as the conceptual schematic, it would have no way of telling /bæad/ from /dæb/ or even /æbd/; all three inputs would simultaneously activate all word nodes containing those three phonemes. What TRACE does is sketched in the rightmost panel of Figure 27.1. Rather than having a single node for the word *bad*, TRACE has many, each aligned with a different point in time. The same scheme applies at the phoneme level: There are /b/ nodes aligned at successive time slices. This provides the model with phonemic and lexical memory—the "trace" behind the model's name.

Consider the input in the center of Figure 27.1. The featural code corresponding to /b/ spreads from time 0 to time 11. There are /b/ nodes aligned with every time slice. So when the input is "heard" by the model, the input at slice 1 is fed to feature nodes aligned with slice 1, then the input at slice 2 is fed to aligned feature nodes, and so on. Feature nodes send activation forward to phoneme nodes aligned with them. Phoneme nodes in turn send activation forward to word nodes aligned with them. In Figure 27.1, connections between /b/ nodes at the right edge of the figure to a *bad* node are shown. Those

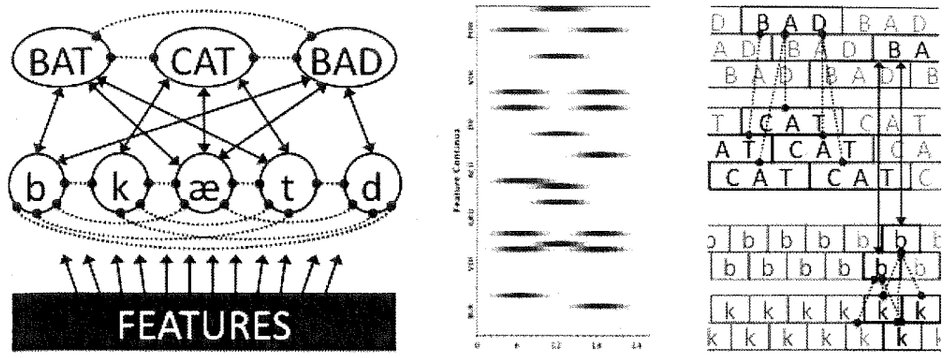

Figure 27.1 The TRACE Model. (*Left*) Conceptual schematic showing feedforward connections from features to phonemes and phonemes to words, feedback from words to phonemes, and lateral inhibition (dashed "bulb" connectors) within phoneme and lexical levels. (*Middle*) TRACE input for the word "bat," showing how activations ramp up at different levels of different features for phonemes, and how phonemes overlap in time. (*Right*) More detailed schematic of the actual design of phoneme and lexical levels in TRACE. Each phoneme node is replicated at multiple time steps, allowing TRACE to "spatialize" time, such that independent nodes can detect the presence of the same or different phonemes at different points in time. The same scheme is used at the word level. Only the phoneme-word connections for one of the last "BAD" word nodes from /b/ are shown. Only the lateral inhibition connections from those /b/ nodes to the /k/ nodes they would inhibit are shown (nodes only inhibit nodes with which they overlap in time). Only the lexical lateral inhibition nodes between one highlighted BAD node and the CAT nodes with which it overlaps in time are shown.

/b/ nodes also have (mutually) inhibitory links with all other phoneme nodes *that overlap with them in time*. This last point is crucial, as it allows multiple phoneme nodes to become highly active; that is, a series of nodes that do not overlap in time can become highly activated since they do not inhibit each other. This allows the model to robustly represent temporal order, since each phoneme is associated not just with a featural code, but with that featural code linked to a particular point in time. For a word node to get robustly activated, the correct temporal series of phonemes must be activated. Just as this scheme allows a series of phonemes to become strongly activated, it also allows TRACE to "recognize" a series of words; nodes aligned with the points in time where the words occurred get activated and do not interfere with earlier and later words with which they do not overlap in time.

A common way for TRACE's behavior to be quantified is with a time course plot, as in Figure 27.2, where we use an example that illustrates how TRACE handles embedded word effects and differences it predicts in the time course of biases for short versus long words. On the left, the input is the word *artist*; on the right, it is *art*. We track activations of several words activated by this input. Note that there is an early short-word advantage apparent in both panels. We also see a late advantage for long words: Note how dramatically higher the activation for *artist* on the left becomes compared to that for *art* on the right. Finally, note as well that TRACE is handling the embedded word problem—*artist* eventually wins the competition on the left despite the fact that the entire patterns corresponding to *are* and *art* have been encountered. The early short-word advantage and late long-word advantage are particularly interesting because Pitt and Samuel (2006) reported evidence for both.

Let's walk through how TRACE does this. The easiest piece of this to understand is the basis for the late advantage for long words: Long words simply accrue more bottom-up activation than short words—the more phonemes there are in a word, the more feedforward activation it will receive. There is a flip side to this advantage, though: Because word nodes receive inhibition from word nodes with which they overlap in time, longer words have more "inhibition sites" than shorter words (that is, they simply overlap temporally with more word nodes because they extend further in time), which puts them at a disadvantage compared to shorter words. Even the tiny bit of bottom-up activation that nodes send when they do not have strong similarity to the input can have a large effect.[8] In the case of embeddings, the fact that *artist* receives more bottom-up input allows it to eventually overcome the activations of embedded words because it can send them more inhibition than they can send to it.

The aspect of TRACE that is conceptually difficult is understanding where the activations in a time course plot come from. This is illustrated in the upper panels of Figure 27.3, which also illustrates

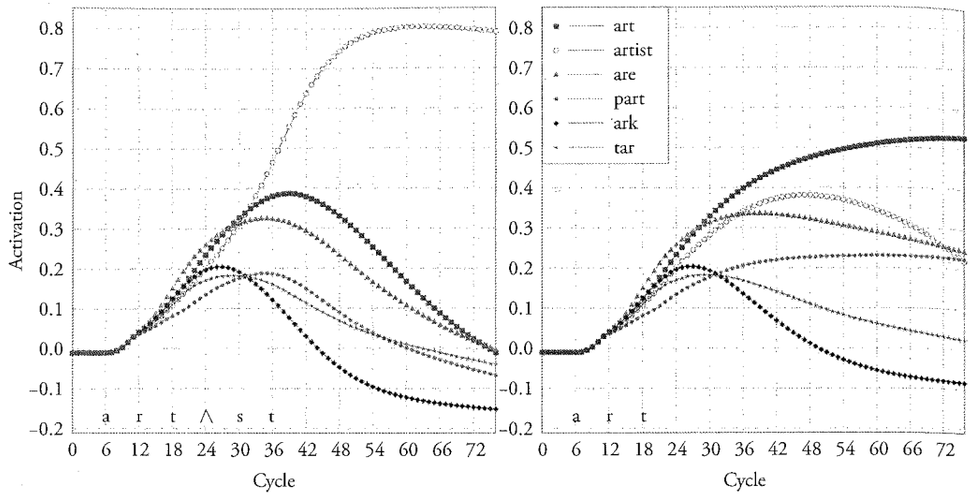

Figure 27.2 Embedded word and word length effects in TRACE. TRACE phonemic inputs are shown in the lower left of each panel, aligned with the cycle at which they were presented. TRACE's handling of embedded words can be seen on the left, where *artist* wins out over even the fully embedded word, *art*. The early short-word bias can also be seen on the left, where the activations of *art* and *are* rise more quickly than that of *artist*. This can also be seen by comparing how quickly *artist* rises in the left panel (hitting activation of 0.3 at cycle 30) and how quickly *art* rises in the right panel (hitting 0.3 around cycle 27). The late advantage for long words can be seen in comparing the peak target activations on the left and right.

what word "segmentation" looks like in TRACE. The figure shows what happens as the series of words *boy pats dog* (/buipatsdag/) has been presented to the network. The "floating" phonemes, for example, have a specific "temporal alignment"—a slice of the model's memory "trace." As time goes by, phonemes (and words) at particular alignments become more or less active as a function of their current and earlier bottom-up support, top-down support (lexical feedback to phonemes), and inhibition from other nodes at the same level. You can also see that the network is successfully inhibiting the word, *pat*, which is embedded in *pats*, and the word *stack*, which is highly similar to the string of phonemes straddling the second word boundary, /sdag/. The bottom panel of Figure 27.3 shows the corresponding lexical time course plot, where the maximally active nodes for a set of words of interest are plotted.

THE TIME COURSE OF COMPETITION

The competition dynamics of TRACE also shed light on the competitor set disagreements between the Cohort and Neighborhood Activation models (reviewed earlier). Cohort predicts that words overlapping at onset will activate one another, even if they are of different length, and that activation of words mismatching at onset will be negligible, even if overall (global) similarity is high. NAM ignores onsets and instead posits that global similarity is what matters, predicting that words will compete if they differ by no more than one phoneme. TRACE predicts something in between the two and offers a resolution to this debate. The left panel of Figure 27.4 shows simulations from Allopenna, Magnuson, and Tanenahus (1998), who examined what TRACE predicts for a target word, *beaker*; a cohort competitor, *beetle*; a rhyme, *speaker*; and an unrelated word, *carriage* (the figure actually averages over several item sets, and an analog to the carrier phrase, "click on the . . . ," was presented prior to the target word; also, to conform to the way Allopenna et al. presented simulations, activations less than zero are plotted as zero). As the input unfolds, the target and cohort activate together since they are consistent with the bottom-up input, /bi/. Once the input begins to favor the target (once the /k/ is presented for *beaker*), the activation of cohort items begins to drop off both because of lesser bottom-up support but also because the target is able to inhibit them. Simultaneously, the input has become more similar to the rhyme (*speaker*), and it becomes much more activated than the unrelated comparison item. However, its peak activation remains substantially below that of cohort items—despite its overall greater similarity to the input (in the example set, the rhyme overlaps in four phonemes with the target, but the cohort only overlaps in two). This is

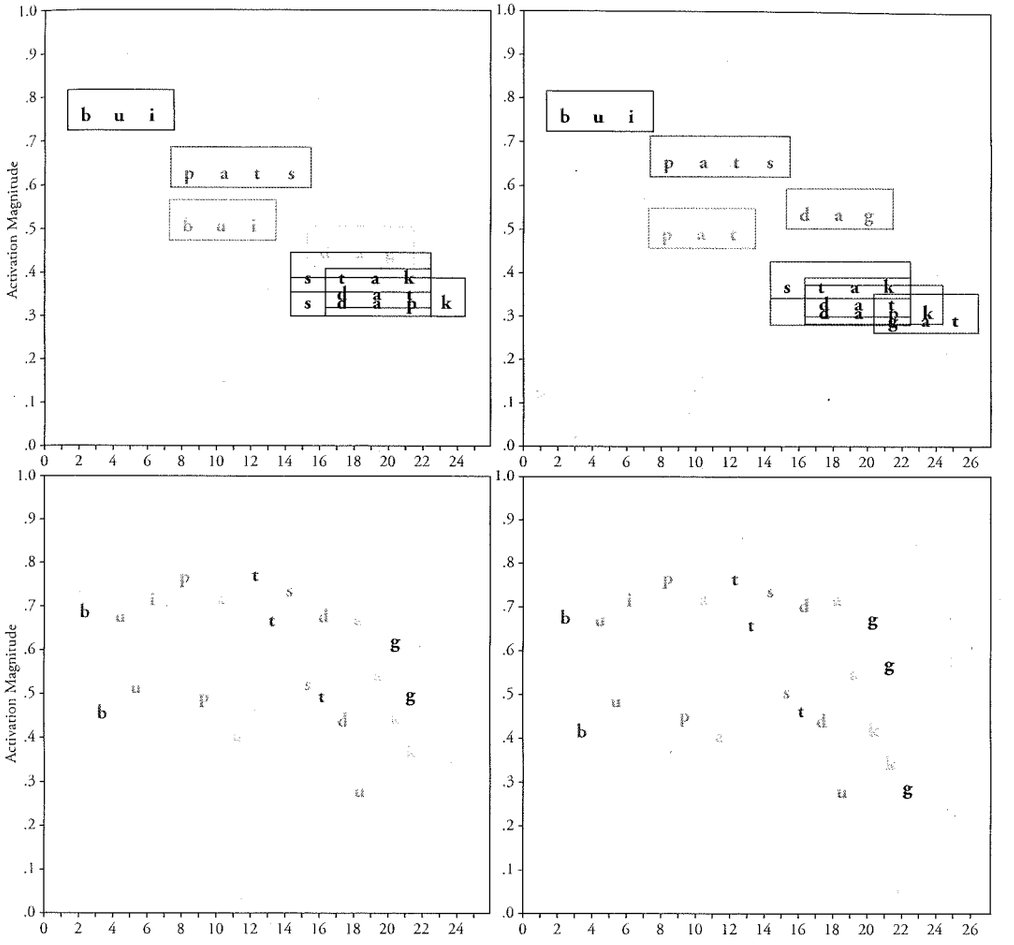

Figure 27.3a Activations over time in TRACE. (*Left*) Snapshots of phoneme (*middle*) and word (*top*) activations just as the /g/ in /buipatsdag/ ("boy pats dog"; note that we are transcribing with phonemes from TRACE's limited inventory rather than accurate English phonemic transcriptions) has been presented (cycle 72, corresponding to temporal alignment 24). (*Right*) Snapshots slightly later (cycle 78, alignment 26). Note the multiple /b/ activations at the left side of the phoneme plots. These show the activations of /b/ nodes aligned with different time slices. The word activations make clear that there is not a magical moment of word recognition in TRACE; rather, there is flux in the relative activation of word units aligned with different portions of the temporal memory "trace." Note, for example, how the DOG (/dag/) node emerges as the late time course "winner" over just the few time steps between the left and right graphs. The plots also illustrate that TRACE solves the lexical embedding problem (PATS wins over PAT). The bottom panel presents a conventional activation time-course plot. Here, each word is represented by the activation of one node—the node for that word that had the highest activation. Note, for example, that activation of *boy* persists, overlapping with high activation of the following words. This does not mean that the model is simulating "hearing" these simultaneously. In the top panel, we can see that the *boy* node we are tracking is aligned with temporal slice 2, which corresponds to processing cycle 6, while the maximally activated (and therefore tracked) node for *pats* aligns with temporal slice 8/processing cycle 24, and the tracked node for *dog* is at slice 16/cycle 48. This illustrates how TRACE maintains an active memory of what words have been presented over time; word activations tracked here are linked to specific instants in memory, such that the tracked activation for *boy* indicates not just that *boy* is active, but that it occurred at a specific time (position, really) in memory. Plots were generated using jTRACE (Strauss, Harris, & Magnuson, 2007).

because it is inhibited not just by the already-activated target but also by the cohort items. Thus, the "head start" that the onset overlap affords to cohort items turns into a significant activation advantage compared to rhymes. But how do these time course predictions correspond to human performance?

The other panels of Figure 27.4 illustrate how Allopenna et al. tested these predictions, using the then new "visual world paradigm" (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995.)[9] Subjects saw displays like the one in the center panel (there were also filler trials where all items were

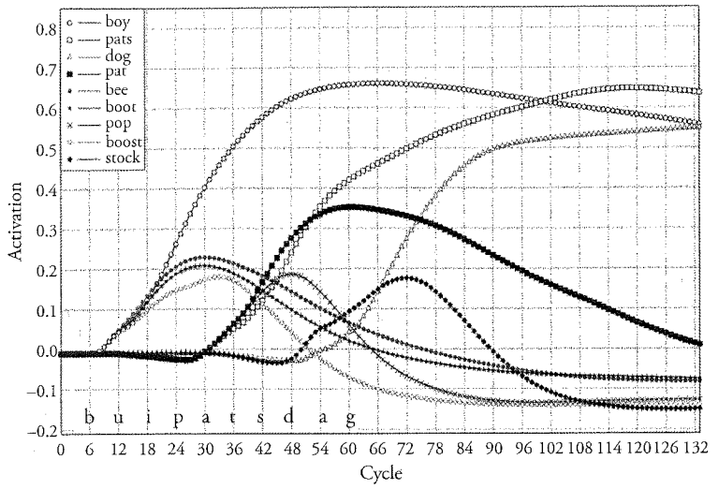

Figure 27.3b (*Continued*)

unrelated). Subjects heard a verbal instruction to pick up one of the items and place it relative to one of the shapes on the screen. Allopenna et al. tracked point of gaze using a head-mounted eye tracker as subjects did this. Starting from the onset of a target word, such as *beaker* in the instruction "pick up the beaker," they plotted the mean proportion of fixations to each of the four items at each time step (right panel, Fig. 27.4), which looked remarkably similar to the TRACE predictions (indeed, see Allopenna et al., 1998 for a simple linking hypothesis that transforms raw activations using a variant of the R. D. Luce [1959] choice rule into response probabilities that are virtually indistinguishable from the observed data). Let's examine some essential details.

First, it's important to understand where the fixation proportions plotted over time come from. On a single trial, "proportions" can only be 1 or 0 for any object at any instant—a subject can only fixate one item at a time. On typical trials, subjects only made ~1.5 fixations in the Allopenna et al. study. So a subject might look at the cohort item 250 ms after word onset, and at the target 350 ms after word onset. For that trial, the data would be 1.0 for the central fixation cross and 0.0 for everything else from word onset to 250 ms, 1.0 for the cohort and 0.0 for everything else from 250–350 ms, and 1.0 for the target and 0.0 for everything else from 350 ms until the end of the trial. Trials for each condition are averaged within participants, and then participant means are averaged to arrive at plots like the one in Figure 27.4. Then, the same sorts of statistical approaches that are used for other kinds of psychological data can be applied (e.g., ANOVA; but see Barr, 2008 and Mirman, Dixon, & Magnuson, 2008, for alternative approaches explicitly designed for assessing change over time). Second, when we look at the time course plots, we can see that changes in mean fixation proportions tended to lag behind phonetic details by about 200 ms. Given that it takes approximately 150 ms to plan and launch an eye movement to a point of light in a darkened room (Fischer, 1992; Matin, Shao, & Boff, 1993; Saslow, 1967) and that inter-saccade intervals are typically in the 200–300 ms range in similar tasks, such as visual search (Viviani, 1990), this lag was nearly as short as could theoretically be.[10] Third, proportions of eye movements over time map onto phonetic similarity over time: Target and cohort looks increased in the early time course, and as the bottom-up input favored the target, fixation proportions to the cohort diminished. This is consistent with results from the gating task. Simultaneously, however, the input had become more similar to the rhyme, and the rhyme fixation proportion eventually exceeded that for the unrelated baseline item. Recall that rhymes are never offered as completions in the gating task, nor are they predicted to compete in the Cohort Model. Why might rhyme effects emerge in this task?

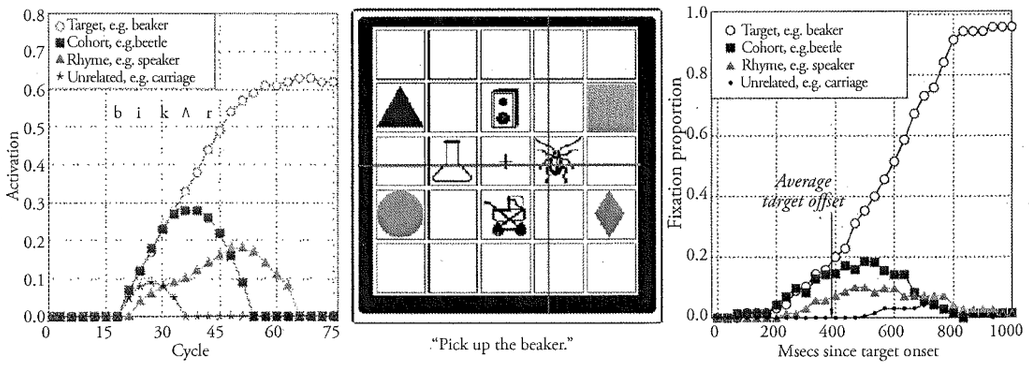

Figure 27.4 (*Left*) Relative activation of cohort (onset) and rhyme competitors in TRACE. (Adapted from Allopenna et al., 1998.) (*Center*) Sample stimulus display and (spoken) instruction from Allopenna et al. (1998). (*Right*) Proportions of fixations over time to each item of interest. Proportions do not sum to 1.0 because fixations outside the four items of interest (e.g., to the central fixation cross) are not plotted.

One possibility is that presenting rhymes in the visual display primes them. Allopenna et al., though, pointed out that fixation proportions map onto phonetic similarity over time—rhymes are not fixated until there is phonetic overlap between them and the input. However, this does not rule out priming; it could be that picture priming boosts the resting level of the rhyme's lexical representation. Allopenna et al. also provided an empirical case for differences between tasks. In a second experiment, they combined the gating task with the visual world task. They presented gated auditory inputs with the same visual displays. Rhyme effects disappeared. Allopenna et al. (1998) argued that this demonstrated that the gating task emphasizes word onsets, leading to overactivation of onset competitors compared to what happens when fluent speech is presented. A later study explicitly tested effects of displaying competitor pictures empirically and with computational modeling (Dahan, Magnuson, Tanenhaus, & Hogan, 2001), and the results suggested that displaying a competitor emphasizes competition effects slightly—but likely because the possibility of fixating the competitor when it is on the screen better *reveals* competition rather than amplifying it. Competition effects persisted when competitors were not displayed. Magnuson, Dixon, Tanenhaus, and Aslin (2007) found effects of frequency-weighted competitor counts (neighborhood and cohort densities, discussed later) even when no competitors were displayed at all.

Now with strong evidence that TRACE simulations of cohort and rhyme effects are largely borne out in time course measures with human subjects, let's return to the debate over the nature of the competitor set. Recall that many results support the Cohort Model (e.g., the absence of evidence for rhyme activation from cross-modal semantic priming), while others support NAM (the variance for recognition and naming data accounted for by the frequency-weighted neighborhood probability metric that includes many offset competitors). Think about what must happen for priming to be detected in the cross-modal semantic priming paradigm. A target like *beaker* must activate a competitor like *beetle* strongly enough that activation spreads to the competitor's semantic associates (like *insect*). If rhymes are activated later and less strongly than cohorts, less activation will spread to their semantic associates, making it difficult to detect priming. By combining the gating and visual world paradigm, Allopenna et al.'s (1998) second experiment suggests that the failure to find evidence of rhyme activation in gating may be an artifact of repeated presentation of onsets in gating, which boosts activation of onset competitors. Thus, it appears that including items that mismatch at onset in the competitor set, as in the NAM, is justified.

However, the NAM is clearly incorrect in excluding onset competitors that mismatch in multiple phonemes (e.g., it would exclude *beetle* from *beaker* neighborhood), and in treating neighbors as equivalent competitors no matter where (when) they differ from a target; clearly, onset competitors need to be weighted more heavily. A study by Magnuson et al. (2007) suggests that the temporal distribution of phonetic similarity has complex impacts on the time course of lexical activation and competition that may be difficult to formalize in a revised NAM metric. Magnuson et al. (2007) attempted to examine the relative contributions of cohort and neighbor definitions by factorially manipulating word frequency, frequency-weighted neighborhood probability, and

frequency-weighted cohort probability. That is, they tested words that were high or low frequency, had low or high neighborhood probability, and low or high cohort probability. Neighborhood probability was estimated by the ratio of a word's log frequency to the summed log frequency of all its neighbors using the NAM 1-phoneme difference definition. Cohort probability was estimated the same way, with cohorts defined as words overlapping in at least the first two phonemes, with no limit on how many subsequent phonemes could differ. Magnuson et al. used the visual world paradigm, presenting target pictures among pictures of three phonologically and semantically unrelated items. They found clear effects of word frequency (high-frequency targets were fixated more quickly than low-frequency targets) and cohort density (low-cohort targets were fixated more quickly than high-cohort targets). The results for neighborhood were more complex. There was an unexpected *early* advantage for high-neighborhood density items, followed by the expected low-neighborhood advantage. This led Magnuson et al. to more carefully examine the makeup of their items' neighborhoods. Although the low- and high-neighborhood items were matched on cohort *density* (that is, the ratio of the target's frequency to the summed frequencies of its cohorts), it turned out that a larger proportion of low-neighborhood items' neighbors were also cohorts. This explained the early disadvantage for low-neighborhood items: Their neighborhoods were "front-loaded." If we consider phonetic overlap between targets and competitors phoneme-by-phoneme, the point of greatest phonetic overlap—and therefore greatest moment of competition—in the low neighborhoods was shifted toward word onset.

THE TIME COURSE OF WORD FREQUENCY EFFECTS

Other experiments comparing the predictions of the TRACE model with time course estimates from the visual world paradigm have revealed that the time course predictions of TRACE are surprisingly robust and very much illustrate the "puzzle solving" of normal science. One example is the temporal locus of frequency effects. Word frequency (frequency of usage) has long been known to influence word recognition, with higher frequency associated with faster and/or more accurate recognition performance (e.g., Howes & Solomon, 1951; Luce & Pisoni, 1998; Marslen-Wilson, 1987). Some cross-modal priming studies suggested that frequency seemed to have an early, transient effect (Marslen-Wilson, 1987; Zwitserlood, 1989). For example, in presenting the word *captain*, one can test for priming of semantic associates of *captain* and its lower frequency onset competitor, *captive*. Stronger priming is found for visually presented associates of higher frequency items (for this example, more priming is found for *ship* than *guard*), but only when the visual probe is presented as the word is being heard. McQueen (1991) also found larger frequency effects in fast responses in a phoneme decision task than in slow responses—again, suggesting frequency has an early influence. Connine, Titone, and Wang (1993) employed an ingenious method for assessing temporal locus from single-response method. They used the Fox Effect (Fox, 1984), in which participants identify the initial consonant of tokens drawn from a continuum between two words (e.g., *best-pest)*. Fox found that if one endpoint has higher frequency, responses shift in its favor. Connine et al. used this method but manipulated whether a block of trials was biased toward high-frequency words, low-frequency words, or was balanced (mixed). In the low-frequency list, for example, ambiguous tokens from low-high frequency pairs (e.g., *pest-best*) were presented, as was the low-frequency endpoint (e.g., *pest*). The prediction was that if frequency were an integral part of the initial activation and perception of a word, this "extrinsic" frequency manipulation (that is, of the probability of low- or high-frequency words being heard) should not matter, and results should be similar whether items were blocked by frequency or mixed. However, they found that extrinsic frequency did matter—so much so that subjects exhibited a bias for low-frequency responses in low-frequency blocks. Their interpretation was that for subjects to be able to modulate attention to word frequency as a function of their extrinsic frequency manipulation, frequency must apply postperceptually, as a decision-stage bias.

Dahan, Magnuson, and Tanenhaus (2001) explored this issue further by testing three implementations of frequency in the TRACE model: (1) making the resting-level activations of words proportional to word frequency, so that a higher frequency word would have a permanent head start; (2) making phoneme-word connections proportional to word frequency (so that the connection from /b/ to *bed* would be stronger than the connection from /b/ to *bench*), consistent with the idea that experience should tune connection strengths; and (3) as a postperceptual effect, where frequency is suddenly applied late in the processing of a word. The resting-level implementation predicts a constant frequency advantage that diminishes as a word is heard and bottom-up input disambiguates between competitors. The postperceptual bias predicts exactly the same thing, except with no frequency effect at all until the "magical moment" where frequency is applied, at which point it becomes identical to the resting-level simulation. The connection strength simulation predicts a constant effect of frequency, but one that is "gated" by the bottom-up input. That is, the connection strength basis for the frequency effect is the multiplication of the bottom-up signal strength by the frequency-weighted connection. When input is weak, the frequency difference is small. As the input becomes stronger, the difference between low- and high-frequency word units becomes much more robust. Then, as with the other implementations, the frequency effect declines as the bottom-up input disambiguates between high- and low-frequency alternatives. When Dahan and colleagues tested these predictions using the visual world paradigm, they found that fixation proportions over time were best modeled by the connection strength implementation: A constant effect of frequency was apparent in the fixation proportions, but it was subtle near word onset, became robust in response to the ambiguous portion of a word (compatible with two alternatives differing in frequency, e.g., *be-* with the alternatives *bed* and *bench*), and then disappeared once the full word had been heard.

As we mentioned before, our focus on the TRACE model is just to illustrate functions that most current theories regard as necessary—which are the functions we have reviewed thus far. These are as follows: a degree of phonetic sensitivity in input representations (/p/ should be more similar to /b/ than to /s/ or /a/), a sublexical level of representation mediating the signal-lexical mapping, sensitivity to prior probability at the word level (frequency), the ability to segment a series of words without explicit word boundary markers, and the ability to handle embedded words (the latter two details are achieved through lateral inhibition at the word level, and through the reduplication of phoneme and word units over time). Next, let's consider details where there continues to be debate.

Points of Disagreement

LATERAL INHIBITION

First, not all models agree that lateral inhibition at the phoneme level is required. The Cohort Model approach favors bottom-up inhibition, for example (Marslen-Wilson & Warren, 1994), and at least one study has argued that inhibition among phoneme units eliminates sensitivity to phonetic detail too quickly to account for behavioral recovery from "lexical garden-paths" (McMurray, Tanenhaus, & Aslin, 2009). Second, it is not clear that direct lateral inhibition between units is the best way to capture competitive dynamics. Gaskell and Marslen-Wilson (1997, 2002) proposed that "emergent inhibition" among overlapping distributed lexical-semantic representations provides a better account (we discuss this account in more detail later).

WORD SEGMENTATION

Even models that agree that lateral inhibition is required at the word level do not buy into the TRACE strategy of reduplicating phoneme and word units to solve the segmentation problem. The Shortlist Model (Norris, 1994) rejects both unit reduplication and feedback. Instead, Shortlist proposes that as words are activated by their bottom-up fit to spoken input, "shortlists" of lateral inhibition networks are created where activated words compete for recognition. This does away with the need for reduplicated units with massive numbers of feedforward, feedback, and lateral inhibition connections (Hannagan, Magnuson, & Grainger, 2012, estimate that a TRACE implementation with a realistically sized lexicon of 20,000 words and a 2-second memory trace would require about 4 million nodes and 80 billion connections). However, Shortlist requires that there be an as-yet unimplemented mechanism that could wire up the needed networks continuously as speech is heard, as well as an as-yet-unimplemented "lookup" network for finding shortlists reduplicated at each phoneme in the input, making the actual savings in terms of numbers of units and connections somewhat unclear. One particularly intriguing aspect of Shortlist is the use of stress information and the *possible-word constraint* (pressure on the model to arrive at parses that result in only a series of words, without leaving residual phonemes that could not form a word according to the phonotactics of English; for example, if *apple* were recognized on hearing *fapple*, this would leave *f* as "residue" that could not form a word, and indeed, listeners have more trouble noticing they have heard *apple* in that case than on hearing an item like *vuffapple*, where parsing *apple* leaves a word-like remainder, *vuff*; Norris, McQueen, Cutler, & Butterfield, 1997).

INTERACTIVITY

There is disagreement about the need for lexical-sublexical feedback, which aficionados of spoken word recognition will recognize as one of the most visible debates in the field. Such feedback is a common feature of interactive activation models, and it is posited to have two beneficial properties: (1) it provides an implicit implementation of probabilistic phonotactics, with common sequences of phonemes receiving boosted activation via resonance from the many words that contain them[11]; and (2) it makes the model robust against noise, whether external (literal noise in the case of speech) or internal. Note that feedback also crucially provides an explanation for top-down effects on speech perception, but these are side effects of a mechanism that make the model more robust.

Norris, McQueen, and Cutler (2000) marshaled a comprehensive theoretical and empirical case against the need for feedback in word recognition. They claimed that logically, feedback could not improve recognition beyond the best performance possible with well-tuned feedforward connections, and that feedback would override bottom-up input to cause hallucinations. They then went on to demonstrate that the majority of top-down effects could be simulated without feedback in a new extension of the Shortlist model that they dubbed *Merge*. In Merge, phonemes feed to lexical nodes, and there is no lexical-phonemic feedback. To simulate, for example, lexical effects on phoneme decisions, Merge posits a set of *postlexical* phoneme decision units that receive input directly from phoneme input nodes and directly from lexical nodes. This allows phonemic and lexical information to be merged—thus accounting for lexical effects on phoneme decisions—but postperceptually, and without contaminating phonemic processing nodes with top-down input (i.e., avoiding hallucinations).

The argument that feedback does not do anything useful in TRACE was based on simulations by Frauenfelder and Peeters (1998) that tested a set of 21 words selected for other simulations. When this analysis was extended to about 900 words, nearly 75% of words were recognized more quickly with feedback on than off (Magnuson, Strauss, & Harris, 2005). Furthermore, Magnuson et al. (2005) tested the performance of the model as noise was added (recall that a primary motivation for feedback is protecting the model against noise effects) and found that feedback substantially preserved accuracy.

The hallucination argument is surprising in light of Figure 7 of the original TRACE paper (McClelland & Elman, 1986), which clearly shows that lexical feedback *modulates* phonemic activation but does not overwhelm bottom-up input (and so, for an input like *shigarette*, lexical feedback would boost activation of /s/, but /ʃ/ would still be much more active than /s/). Mirman, McClelland, and Holt (2005) explored this issue more thoroughly, confirming that lexical feedback in TRACE could not overwhelm bottom-up input, but it could delay recognition of lexically inconsistent phonemes in some cases. Their behavioral experiments then confirmed that human listeners do indeed exhibit lexically induced delays in phoneme recognition in precisely the contexts predicted by TRACE. Furthermore, Mirman, Bolger, Khaitan, and McClelland (in press) demonstrated that it is trivially easy to balance feedforward and feedback gain to preclude hallucination, and they extended earlier arguments by Movellan and McClelland (2001) that interactive models such as TRACE implement optimal perceptual inference.

Most damaging for the Merge account, there appear to be "knock-on" or "indirect" effects of lexical activation on sublexical processing, which are only possible if lexical information is feeding back directly to sublexical levels during online processing. The first demonstration of such effects was lexically mediated compensation for coarticulation (Elman & McClelland, 1988; see also Magnuson, McMurray, Tanenhaus, & Aslin, 2003a; Samuel & Pitt, 2003), which has been disputed (McQueen, Jesse, & Norris, 2009[12]; Pitt & McQueen, 1998; see also the exchange between McQueen, 2003, and Magnuson, McMurray, Tanenhaus, & Aslin, 2003b). However, there are at least two other such indirect effects: lexically induced selective adaptation (Samuel, 1997, 2001) and lexically guided tuning of speech perception (Norris, McQueen, & Cutler, 2003; for a review see Samuel & Kraljic, 2009). The latter of these is particularly important because it is less controversial (both camps agree it requires feedback, but they debate when feedback happens; Norris et al. argue that the learning signal is somehow stored in memory and applied later—"feedback for learning" rather than the online feedback in TRACE) and has opened new avenues for investigating the representation of speech sounds and the interplay between lexical and sublexical representations (more on this later).

For more detailed discussion of interactivity in speech perception and other cognitive and perceptual domains, see McClelland, Mirman, and Holt (2006), the exchange between McQueen, Norris, and Cutler (2006) and Mirman, McCelland, and Holt (2006), Mirman (2008), and Mirman et al. (in press). While our position on this topic is likely clear, we think it is fair to say that the preponderance of evidence supports interactive processing as a central principle across cognitive and perceptual domains, including spoken word recognition and speech perception. Convincing all parties will likely take substantially more empirical and computational work.

Moving Beyond the Limitations of the TRACE Model

It is essential that we keep TRACE's limitations in mind. TRACE achieves deep and broad coverage only of sound form recognition—not meaning—and it is a "hand-wired" model with fixed parameters set by the experimenter. Given that the goal of spoken word recognition is to support perception of a speaker's message, the absence of semantics and connections to sentence processing in TRACE and nearly all other models of spoken word recognition is a serious gap. And while an accurate model of average adult performance in word recognition is an invaluable tool, it can only be a waypoint in the quest to understand language comprehension, which eventually must include an account of development.

A handful of efforts have been made at addressing spoken word recognition in a developmental context with models of word learning using simple recurrent networks (Christiansen, Allen, & Seidenberg, 1998; Gaskell & Marslen-Wilson, 1997, 1999; Magnuson, Tanenhaus, Aslin, & Dahan, 2003). Each of these has modeled some important aspects of word learning and word recognition, but the Gaskell and Marslen-Wilson *Distributed Cohort Model* stands out for tackling form and meaning simultaneously. The Distributed Cohort Model simulates several aspects of spoken word recognition (largely consistent with the earlier Cohort Model) and leads to some new puzzles for models of spoken word recognition to grapple with that would have been difficult to intuit without a working model that simultaneously activates form and meaning. For example, in the midst of phonological competition (e.g., between *captain* and *captive* when just the /kæpt/ portion has been heard), semantic representations will be in a rather odd state: a blend of the semantic features of the items in the phonologically activated competitor set. This has the potential to interact with phonological competition and to provide insights into how context might influence form recognition.

This is an important consideration, as during the current normal science phase of spoken word recognition research, a handful of visual world paradigm studies on semantic competition and contextual constraints have appeared that are at odds with the older research we reviewed earlier—and with each other. First, consider evidence of semantic competition. At the coarsest level, visual world studies show that semantic competitors include words that are members of the same semantic category (e.g., *piano—trumpet*, Huettig & Altmann, 2005), semantic associates (e.g., *ham—eggs*, Yee & Sedivy, 2006), and concepts that frequently co-occur in situations or events (e.g., *balloon—clown*, Mirman & Graziano, 2011). Just as greater phonological similarity produces stronger phonological competition, greater semantic similarity produces more semantic competition (Mirman & Magnuson, 2009). Importantly, these competition effects are truly "semantic" in nature—simple lexical co-occurrence is not sufficient to cause the effect: When hearing *lettuce*, there is no evidence of activation of *iceberg*, even though the two words frequently co-occur (Yee, Overton, & Thompson-Schill, 2009). A concern with studying semantically related items in the visual world paradigm is that competition may be induced just by presenting pictures of related items. A follow-up study conducted by Yee and Sedivy (2006) addresses this possibility. Again, they found semantic competition effects that resembled (in terms of timing and magnitude) the phonological competition effects seen in Figure 27.4. For example, as subjects heard *lock*, they were significantly more likely to fixate a picture of a *key* than an unrelated item. In a second experiment, they tested whether this could be due just to picture priming by looking for phonologically mediated semantic activation. Instead of using *lock* as the target, they used *logs*. The logic was that as the participant hears *logs*, *lock* should be activated by phonological similarity, and then should spread activation to *key* via semantic associations. This is just what they found.[13]

Semantic competition can be driven by specific semantic features. For example, there is partial activation of objects that are similar in shape (e.g., *rope—snake*, Dahan & Tanenhaus, 2005; Yee, Huffstetler, & Thompson-Schill, 2011), motor actions (e.g., *piano—typewriter*, Lee, Middleton, Mirman, Kalenine, & Buxbaum, in press; Myung, Blumstein, & Sedivy, 2006), or function (e.g., *broom—sponge*, Kalenine, Mirman, Middeton, & Buxbaum, in press; Yee et al., 2011). In addition, there appears to be a time course to semantic feature

activation: Kalenine et al. found that features shared by objects that are used together (e.g., *broom—dustpan*) are activated faster than general function features (e.g., *broom—sponge*), Lee et al. found that structural action features are activated faster than functional action features (i.e., "grasp" features faster than "use" features), and shape features may be activated faster than function features (Yee et al., 2011).

The fact that people look to items with only tangential connection to the bottom-up input presents us with yet another puzzle when we consider visual world studies looking at sentential and other constraints, which suggest that lexical competition can be flexibly restricted. Dahan and Tanenhaus (2004) used the visual world and found that sentential contexts in Dutch like "never before climbed a..." led to a strong and even anticipatory advantage for the Dutch word for *goat* (*bok*) compared to a phonological competitor (*bot*, Dutch for *bone*). A manipulation of the speech file favoring *bot* was able to modulate the result, demonstrating the online interplay between sentential constraints and details of the speech signal. Chambers, Tanenhaus, and Magnuson (2004) had subjects follow instructions to interact with real objects. When subjects were told to pick up an *egg* and there were two eggs available—one in the shell and a liquid one that had been cracked into a bowl—subjects looked at both and needed clarification to proceed. But if the instruction were to *pour the egg*, subjects did not even look at the unpourable egg still in the shell (similar results were found as a function of the affordances of implements; subjects holding a hook restricted attention to "hookable" objects when they were told to pick up some object!). Magnuson, Tanenhaus, and Aslin (2008) used an artificial lexicon paradigm to implement a stronger version of the Tanenhaus et al. (1979) and Shillcock and Bard (1993) experiments we reviewed earlier and found that strong form-class expectations based on the pragmatics of visual displays (whether bare noun reference would suffice, or whether an adjective was required for unambiguous reference) were able to wipe out cross-form class competition (phonologically similar nouns and adjectives did not compete with each other). Similarly, pragmatic expectations about which objects will be referred to by a speaker appear to influence lexical activation (Barr, 2008; Hanna, Tanenhaus, & Trueswell, 2003). The puzzle here is that on the one hand, the semantic competition effects suggest promiscuous, automatic spreading of activation over pathways sensitive to phonological and semantic relatedness, but at the same time, strong sentential and pragmatic constraints seem able to prevent activation of individual items or even classes of items. At first blush, the semantic effects appear to be a case of facilitation, but the sentential/pragmatic constraint effects appear to be inhibition. It may be that both kinds of findings have to do with facilitation; sentential/pragmatic constraints may just boost activation of compatible items (see footnote 12). However, it is difficult to intuit how such mechanisms would operate without implementing them in a simulation model, and it appears that progress on this front will require implemented models.

Similar findings are emerging at the interface of words and sentences. There is growing evidence that words activate in an anticipatory fashion based on context and inferences about what will be heard from properties of animate and inanimate objects depicted in a visual scene (e.g., Altmann & Kamide, 1999; see Altmann & Mirkovic, 2009, for a review, and Kukona, Fang, Aicher, Chen, & Magnuson, 2011, for details about the time course of sentential and lexical-thematic constraints). As with the priming literature, understanding whether such mechanisms are driven primarily by facilitation or inhibition, and how lexical and sentential constraints interact, will require implemented models.

On our view, we are in the thick of the normal science phase of research on spoken word recognition. The field finds itself with a rough consensus about the data that need to be explained, the computational theory of the input-output mapping, and the essential functions the mechanisms of spoken word recognition must include. There is vastly more agreement than disagreement about these matters. However, a closer look at the data we have just reviewed reveals some discomfiting incongruencies between the empirical findings and our current models. First, no current model can accommodate *all* of the results we have just reviewed; TRACE (McClelland & Elman, 1986) simulates a surprisingly broad and deep array of form recognition phenomena at a fine-grained time scale, but it cannot address meaning. The Distributed Cohort Model (Gaskell & Marslen-Wilson, 1997, 1999) is a learning model that simulates a combination of form and meaning phenomena and reveals interesting problems that emerge when such representations interact. However, it is unknown whether that model can be extended to the range of phenomena TRACE simulates. Furthermore, there are some results that extant models cannot accommodate.

One simple example is priming, whether phonological, phonetic, or semantic. Priming seems like it should be easy to accommodate in cognitive psychological models of language processing; it is a basic, long-standing phenomenon that has motivated and constrained theories in many cognitive domains—how could priming be incompatible with current theories of spoken word recognition? All the same, most current theories of spoken word recognition cannot accommodate it (we will discuss the one counterexample, Goldinger [1998], later). Claims that priming effects can be modeled by connectionist models of spoken word recognition (Luce et al., 2000) typically boil down to the fact that simultaneous activations of prime, probe, and baseline items (e.g., *veer, bull, soft,* where *veer* primes *bull* due to high phonetic similarity at each phoneme) within the model show a pattern of some sort of connection (e.g., simply that the prime and probe are both active). What current simulation models are incapable of is showing priming like that of human listeners, where the prime has a postive or negative impact on *subsequent* processing of the probe. This may not seem like a crucial example, but it illustrates the fact that our theories and models may do an even worse job of accounting for empirical facts than we realize.

This leaves us with a consensus but without a unified model. Instead, we have separate microtheories of form and meaning, activation and priming, and so on, that nonetheless appear to be compatible with one another in broad strokes because each is consistent with general principles of cognitive psychology. On top of these gaps, though, there are more troubling anomalies looming—results that seem truly incompatible with current theories and that may require a new theoretical formulation.

Waiting for the Revolution

The gaps we have just reviewed are minor in comparison to a growing set of results that are truly incommensurate with current theories. Kuhn (1962) argues that scientific revolutions, or paradigm shifts, are triggered by the accrual of so many anomalous results that a dominant paradigm cannot continue in the normal science mode of incorporating more and more fine-grained details into an existing theoretical framework. Normal science can withstand a number of anomalies; they can be ignored, treated as curiosities, or so nearly compatible with current theory that integration seems imminent. Eventually, though, so many anomalies accrue that a tipping point is reached. It would appear that we are nearing this point in spoken word recognition. In this section we present a selective review of the most pressing anomalies.

Surface Details in the Speech Signal

We began this chapter with a discussion of the simplifying assumption that the input to spoken word recognition can be temporarily assumed to be a string of phonemes output by a speech perception mechanism. Such simplifying assumptions are common throughout science and allow progress at multiple levels of analysis, rather than a purely bottom-up approach that does not progress until all fundamental problems are cracked. However, temporary simplifying assumptions can take on the functional status of true theoretical assumptions and hide constraints—ironically becoming complicating assumptions (Magnuson, 2008). Let's consider how the phonemic input assumption has done just that.

A fundamental finding in speech perception is the categorical perception of many speech sounds (Liberman, Harris, Hoffman, & Griffith, 1957). There tends to be a sharp boundary between phonemic categories, such that a change in a critical dimension (e.g., voice onset time) is difficult to detect within a category, but change of the same magnitude that straddles the category boundary is obvious. However, it has been known for a long time that human speech perception *is* sensitive to within-category variation (e.g., Pisoni & Tash, 1974) and that that sensitivity can be observed in spoken word recognition (Andruski, Blumstein, & Burton, 1994; Marslen-Wilson & Warren, 1994). Studies using the visual world paradigm have demonstrated this sensitivity in great detail in speech perception (e.g., McMurray, Tanenhaus, & Aslin, 2002), as well as exquisite sensitivity to subcategorical (subphonemic) detail, such as coarticulatory cues (Dahan, Magnuson, Tanenhaus, & Hogan, 2001). These results should have led directly to the insight that the phonemic input assumption is just plain wrong, and that the computational theory of spoken word recognition cannot be compartmentalized away from speech perception—especially in light of well-known phonetic studies demonstrating, for example, that vowel durations differ systematically (albeit probabilistically) as a function of word length (e.g., Peterson & Lehiste, 1960). Salverda, Dahan, and McQueen (2003) finally made this connection explicit in a visual world study (see also Davis, Marslen-Wilson, & Gaskell, 2002, for related results using a priming paradigm).

They confirmed that, on average, vowel durations in longer words (e.g., *hamster*) were shorter (by only about 20 ms at an average speaking rate) than in shorter words (e.g., *ham*), and then demonstrated that listeners use subtle cues like these immediately to constrain spoken word recognition (e.g., the *ham-* of *hamster* leads to greater activation of *hammer*, which is compatible with cues to word length present already in /æ/, than the short word, *ham*). On the one hand, building in a correlation between word length and vowel duration would be fairly easy in a model like TRACE; on the other hand, such durations just scratch the surface of the complex prosodic patterns to which listeners are sensitive (see Salverda et al. for some of these). Consider the implications for the *embedding problem* if the initial segments in spoken words essentially tell the listener how long the word they are hearing will be; this substantially mitigates the embedding problem (i.e., if vowel duration is consistent with a two-syllable word, *ham* should be less active, reducing the magnitude of the embedding problem). Thus, simplifying assumptions about the nature of the input hid subphonemic constraints available in the speech signal.

This is just one example of surface details that spoken word recognition is sensitive to. Another is assimilation. In English, *place assimilation* can straddle word boundaries, with actual place of articulation shifting with the context of an adjacent segment. For example, the sequence *green boat* is often pronounced as a rhyme or near rhyme of *dream boat*, but this appears not to impede perception of the intended word. If there is truly full assimilation (e.g., Gaskell, 2003; Gaskell & Marslen-Wilson, 1998), this might be a problem situated at the level of spoken word recognition. However, evidence that characteristics of the intended segment shift only partially toward the assimilating place (e.g., Gow, 2001, 2002, 2003a, 2003b) suggest a degree of interaction between speech perception and spoken word recognition that cannot be trivially accommodated in current models, especially when the phonemic input assumption is in play.

Word segmentation is another place where the phonemic input assumption may be complicating rather than simplifying. Relatively easy to quantify surface details—such as stress patterns—are not difficult to integrate with current frameworks (Norris, McQueen, & Cutler, 1995). Mattys and colleagues (e.g., Mattys, White, & Melhorn, 2005) have been mapping out complex interactions between knowledge-based (lexical) and signal-based cues to segmentation, the relative weights of which vary with context. Again, aspects of these details could be built into a model like TRACE. But it does not seem plausible that the full number and complexity of constraints Mattys and colleagues have documented could be hand-coded into such models. In fact, attempting to do so would amount to replacing the phonemic input assumption with another simplifying assumption, albeit one that would be more complex and realistic. That is, building in such correlations with any simplified/abstract analog to the acoustics of real speech is still a matter of constructing a pretend signal, rather than grappling with actual speech. Without tackling the signal, we will not know what helpful constraints we have hidden with the abstractions of our simplifying assumptions.

A case where surface detail seems complicating rather than simplifying, though, is the great sensitivity listeners exhibit to surface details that do not seem to provide general constraints on the signal-word mapping reported by Goldinger (1998). Memory for and naming of a word heard earlier in a session (e.g., in an old-new recognition test) benefit from preservation of seemingly nonessential details such as talker identity. Goldinger (1998) proposed an episodic model of the lexicon, where the storage of unanalyzed memory traces provides the basis for word recognition. While several fundamental problems would have to be overcome for this to be a viable model of word recognition (see Magnuson & Nusbaum, 2007, and footnote 2), Goldinger's simulations with an episodic model (based on MINERVA2; Hintzman, 1986) provide a potential starting point for understanding priming and recognition within a single system.

Further need for some sort of flexible, active, and context-dependent episodic learning based on complex contingencies between talkers and surface details is illustrated by recent studies showing rapid, but conservative, learning. Norris, McQueen, and Cutler (2003) and Bertelson, Vroomen, and DeGelder (2003) reported rapid "perceptual recalibration" when ambiguous phonetic stimuli are accompanied by some kind of disambiguating context. Bertelson et al. used videos of the oral gestures of speakers to disambiguate segments halfway between /b/ and /d/. Norris et al. predicted that lexically disambiguated segments (Ganong, 1980) could lead to lexically guided learning. They used segments halfway between /f/ and /s/ that were lexically disambiguated as subjects did a lexical decision task (e.g., in English, the ambiguous token in the context of *distre-* would most likely be *distress*, but the same ambiguous segment would most likely be /f/ in the context of *himsel-*). In both studies, the contexts drove learning that changed performance on subsequent phonetic identification assessments. Thus, phonetic-to-phonemic mappings are not static but can change dynamically based on recent experience. Later studies showed that such learning is robust over time (25 minutes: Kraljic & Samuel, 2005; 12 hours: Eisner & McQueen, 2006) and re-exposure to canonical, unambiguous tokens (Kraljic & Samuel, 2005). Perhaps most remarkably, this learning appears to depend on history with a speaker *and* the absence of a causal explanation for perturbed speech production; Kraljic, Samuel, and Brennan (2008) found that learning is talker-specific and depends on the first utterances heard by a particular talker (early unambiguous tokens are not overridden by later, lexically disambiguated tokens) *and* that such learning is blocked if a causal, external factor would explain temporary deviance from canonical pronunciation (such as a pen in the talker's mouth).

One last related phenomenon we will touch on here is word learning in adults, which both highlights further the need to grapple with learning and the need to integrate the cognitive psychology of language with the cognitive neuroscience of language (and eventually the neurochemistry and genetics, etc.). Gaskell and Dumay (2003) pioneered a new learning paradigm where subjects are exposed to word-like patterns (e.g., *cathedruke*) as nonword foils in lexical decision over a period of 5 days. They found that eventually, these nonwords began to act as though they were becoming lexicalized, as indicated by increased competition evident when subjects processed similar real words (e.g., *cathedral*). Dumay and Gaskell (2007) explicitly tested whether the advent of lexical competition is dependent upon sleep-based consolidation. With a classic sleep + delay versus no-sleep + delay approach, they found that the emergence of competition does depend on consolidation. Leach and Samuel (2007) pointed out that this requires two different phases of learning, and possibly two different forms of learning: *lexical configuration* is a matter of learning to recognize a pattern, and this seems to emerge before *lexical engagement*, when that pattern begins to show evidence (via competition with existing lexical items) of integration with the lexicon. While in this chapter we have steadfastly stuck to our charge of reviewing the cognitive psychology of spoken word recognition, ambitious behavioral and neuroimaging work on consolidation in lexical learning by Davis and Gaskell (2009) at the intersection of neuroscience and cognitive psychology illustrates that maintaining boundaries between these domains is becoming less tenable every day. We only have time to mention in passing the growing need for theories and models of spoken word recognition to respect computational constraints emerging from the cognitive neuroscience of language, and the fact that this represents another source of Kuhnian crisis for current theories.

Revolutionary Frameworks?

In closing, we see three avenues as most promising for pushing the field beyond the tipping point and to new theoretical frameworks: (1) the need to grapple with the speech signal itself, (2) integration of the study of spoken word recognition with descriptively higher levels of language processing, (3) the need for theories and models to grapple with learning across the life span, including language development in childhood and rapid, flexible learning in adults, and (4) the need to respect neurobiological constraints on mechanisms for language processing. Two existing approaches might provide paths forward on some of these.

The first is the adaptive resonance framework of Grossberg and colleagues (e.g., Grossberg, Boardman, & Cohen, 1997; Grossberg, Govindarajan, Wyse, & Cohen, 2004; Grossberg & Myers, 2000), who have ignored the simplifying assumptions embraced elsewhere in the field and have stubbornly refused to abandon the speech signal or neurobiologically realistic learning models. A pessimistic view of this work would be that progress has been slight, leading thus far neither to anything like the breadth of the TRACE model nor to a plausible developmental model. As we have discussed, simplifying assumptions about the input signal can actually complicate the problem of spoken word recognition. Working with the signal is necessary both in terms of demonstrating that our models could actually work with the signal whose perception we intend to model, and in uncovering further constraints our simplifying assumptions have hidden. Similar benefits may also emerge from realistic neural modeling.

The second promising framework is a connectionist model that integrates the development of speech perception and speech production (Plaut & Kello, 1999), using acoustic and articulatory representations that are still abstractions but are tremendously more realistic than TRACE (McClelland & Elman, 1986), the next best simplified approach in current models, providing a waypoint perhaps between

current approaches and adaptive resonance. The model learns to recognize words based on "adult" input and learns to control its articulatory apparatus by attempting to mimic the sound patterns in that input. More generally, it may be time for theories and models of spoken word recognition to move away from stipulated representations to emergent representations (McClelland, 2010; McClelland et al., 2010).

But, of course, a framework encompassing learning is even more important when we turn our attention to language development. As we suggested earlier, models of adult "endpoints" are invaluable tools for making progress on understanding spoken word recognition (even if endpoints are another example of a useful but misleading simplifying assumption, given the results for adult plasticity we have just reviewed), but full understanding of language processing will require developmental models. Plaut and Kello did not develop the model beyond this first report, but a framework of this sort may be just what is needed to push the field forward.

Notes

1. Typical speaking rates range from approximately 12 phonemes per second (moderate) to 30 (fast) phonemes per second (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Even though 12 phonemes per second is already three to four times the rate at which sequences of arbitrary sounds can be perceived (Warren, Obusek, Farmer, & Warren, 1969), word recognition remains stable at unusually fast rates up to 50–60 phonemes per second.

2. Although the field continues to be known as "spoken word *recognition,*" we will see that current theories largely eschew the notion of a "magical moment" (Balota, 1990) where the system enters a state of one word being recognized, embracing instead notions of change over time in relative activations of lexical representations in memory (e.g., McClelland & Elman, 1986) or Bayesian probability estimates that particular words have been or are being heard (e.g., Norris & McQueen, 2008). For better or worse, we will follow standard practice of using "spoken word recognition" as a catchall term to indicate processes intervening between speech perception and sentence processing.

3. Whether one must fully commit to abstract or episodic representations is a complex issue beyond the scope of this chapter. See Magnuson and Nusbaum (2007) for a discussion of the possibility that both support speech perception and word recognition.

4. Note that a more conventionalway of reviewing spoken word recognition is to divide the problem is divided into three functions: access, selection, and integration (e.g., Dahan & Magnuson, 2006; Frauenfelder & Tyler, 1987; Marslen-Wilson, 1987). Access, or initial contact, is how the speech signal is mapped to phonological (or possibly other) representational forms that are the key to accessing lexical entries. The selection function identifies the lexical item with the best fit to the bottom-up input, possibly constrained by top-down context. The integration function must output a form that can be the basis for the syntactic and semantic processing. While we agree these are necessary functions, we do not view them as necessarily independent, nor even distinct.

5. Or rather, many of the basic facts reemerged. Bagley (1900) reports the results of a research program that presaged much of the basic work at the intersection of speech perception and spoken word recognition of the 1970s and 1980s, including ingenious modifications ("mutilations") of wax drum recordings for manipulation of spoken words. In the interest of concision and utility, we will focus on recent work with modern techniques, even when Bagley's work presents an interesting precedent. Intrigued readers should see the comparison of Bagley's methods and results with those of 1970s and 1980s psycholinguists by Cole and Rudnick (1983).

6. We must mention two things here. First, these latter two problems (resolving phonological patterns and segmenting phonemes) are problems typically deferred to speech perception specialists, via the simplifying assumption that the input is a series of parsed phonemes, as reviewed earlier. Second, it is not a given that the speech signal must be parsed into phonemes before it can be mapped onto words, even though almost all theories of spoken word recognition assume there is some level of sublexical representation mediating the mapping from acoustics to words; for example, Klatt (1979) proposed a theory in which acoustic patterns (spectra) were mapped directly to words.

7. Note that this avoids the problems with interaction that led to the abandonment of top-down interaction in the Cohort Model (Marslen-Wilson, 1987, reviewed above). Even without interaction, the notion of the recognition point in the Cohort Model has serious problems. We have no trouble realizing we have *not* just heard a word when we hear *banisfer*, even though we hit the uniqueness point for *banister* at the /s/—pushing Cohort to predict recognition of *banister*. If we view this as a case where a word has reached a high degree of activation relative to other words, rather than reaching an all-or-none recognition state, we avoid this problem. High activation allows a listener to *expect* a word and to perform as though it has been recognized, but without absolute commitment to that word.

8. The interested reader can test this for herself using jTRACE (Strauss et al., 2007); even if you reduce the lexicon to a single pair, such as *artist* and *art*, there is still an early short-word advantage because each *artist* node receives more inhibition from overlapping *art* nodes, and the only way to eliminate the effect completely is to turn off lateral inhibition. A short-word advantage persists even if you reduce the lexicon to one long word and one unrelated short word (e.g., *artist* and *blue*) and compare the recognition time for the two words (again, you can eliminate the bias by turning off lateral inhibition).

9. An early version of this task was reported by Cooper (1974), but its potential was not appreciated at the time. Recently, a version was developed that tracks hand movements instead of eye movements and provides a different perspective on time course (Spivey, Grosjean, & Knoblish, 2005).

10. Note that 200 ms is the mean lag, which means that there are faster and slower initial fixations. See Altmann and Kamide (2007; also Altmann, 2011) for assessments of saccade latency distributions.

11. We are using the term *probabilistic phonotactics* here to capture the general idea that words can vary in just how word-like they sound, with more common sequences sounding more word-like. This term also has been given an operational definition in spoken word recognition research as well. Vitevitch and Luce (1999) define this as a sum of diphone probabilities within a word (so not exactly a probability). While this is strongly correlated with neighborhood density, Vitevitch and Luce have demonstrated that in a task where attention is manipulated to a sublexical locus (by presenting many nonwords in a naming task, for example), phonotactic probability facilitates processing.

12. Indeed, McQueen et al. (2010) failed to replicate the findings of Magnuson et al. (2003a) using the original materials. Using better materials provided by Magnuson et al., McQueen et al. mostly found null effects, as well as one reliable effect in the same direction as Magnuson et al. and one in the opposite direction, which they argue depended on perceptual learning based on the types of items included in practice trials. While positive results still outnumber negative ones (see Magnuson et al., 2003b, for a scorecard), it is clear that this experimental paradigm is fragile, and that resolving this debate may require a different form of evidence.

13. There is a puzzle lurking here. After seeing or hearing one item and then responding to a semantically related item, we typically observe classical positive priming: lexical decision or naming responses to the second item are speeded relative to responses to that item when preceded by an unrelated item. But inhibitory priming is observed for phonological relatedness in such tasks (Luce et al., 2000). The puzzle emerges when we consider what happens in the visual world paradigm. The presence of phonologically related items results in apparent competition; fixations are diverted to the competitor, and target fixations are proportionally depressed. This is interpreted both as evidence for activation of the competitor and inhibition of the target. When a semantic associate is present, the same pattern is observed. However, this may be a case where the paradigm masks the true effect; if the semantic associate is activated, it will attract fixations by virtue of that activation, but it may not actively inhibit activation of the target. This might also suggest the caution is warranted in interpreting the phonological case as demonstrating both competitor activation and target inhibition. All the same, we shall follow the literature and refer to both phonological and semantic cases as instances of competition.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439.

Altman, G. T. M. (2010). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica, 137*(2), 190–200.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247–264.

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57*, 502–518.

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33*(4), 583–609.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition, 52*, 163–187.

Aslin, R. N., Woodward, J., LaMendola, N., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Mahwah, NJ: Erlbaum.

Balota, D. A. (1990). The role of meaning in word recogni- tion. In D. A Balota, G. B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 9–32). Hillsdale, NJ: Erlbaum.

Barr, D. J. (2008). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition, 109*(1), 18–40.

Bertelson, P., Vroomen, J., & DeGelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science, 14*(6), 592–597.

Blumstein, S. E. & Myers, E. (in press). Speech perception. In *Oxford handbook of cognitive neuroscience.*

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 687–696.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13*, 221–268.

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.

Connine, C. M., Titone, D., & Wang, J. (1993). Auditory word recognition: Extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology: Learning Memory and Cognition, 19*(1), 81–94.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. A new methodology for the real time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*, 84–107.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42*, 317–367.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Tracking the time course of subcategorical mismatches: Evidence for lexical competition. *Language and Cognitive Processes, 16*(5/6), 507–534.

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 498–513

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review, 12*, 453–459.

Davis, M. H., Marslen-Wilson, W. D., & Gaskell. M. G. (2002). Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 218–244.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society B, 364*, 3773–3800.

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science, 21*(11), 1664–1670.

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science, 18*, 35–39.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America, 119*(4), 1950–1953.

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.

Fischer, B. (1992). Saccadic reaction time: Implications for reading, dyslexia and visual cognition. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 31–45). New York: Springer-Verlag.

Fowler, C. A., & Magnuson, J. S. (2012). Speech perception. In M. Spivey, M. Joanisse, & K. McRae (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 3–25). Cambridge: Cambridge University Press.

Fox, R. A. (1984). Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 526–540.

Frauenfelder, U. H., & Peeters, G. (1998). Simulating the time course of spoken word recognition: An analysis of lexical competition in TRACE. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 101–146). Mahwah, NJ: Erlbaum.

Frauenfelder, U. H., & Tyler, L. K. (1987). The process of spoken word recognition: An introduction. *Cognition, 25*, 1–20.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 110–125.

Gaskell, M. G. (2003). Modelling regressive and progressive effects of assimilation in speech perception. *Journal of Phonetics, 31*, 447–463.

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition, 89*, 105–132.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*(5), 613–656.

Gaskell, M. G., & Marslen-Wilson, W. D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception & Performance, 24*, 380–396.

Gaskell, M. G., & Marslen-Wilson, W. D. (1999). Ambiguity, competition, and blending in spoken word recognition. *Cognitive Science, 23*, 439–462.

Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology, 45*(2), 220–266.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251–279.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language, 28*, 501–518.

Gow, D. W., Jr. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language, 45*, 133–159.

Gow, D. W., Jr. (2002). Does English coronal place assimilation create lexical ambiguity? *Journal of Experimental Psychology: Human Perception and Performance, 28*, 163–179.

Gow, D. W., Jr. (2003a). Feature parsing: Feature cue mapping in spoken word recognition. *Perception and Psychophysics, 65*, 575–590.

Gow, D. W., Jr. (2003b). How representations help define computational problems: Commentary on Grossberg, Gaskell and Greenberg. *Journal of Phonetics, 31*, 487–493.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing visual word recognition: A multiple read-out model. *Psychological Review, 103*, 674–691.

Greenberg, J. H., & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word, 20*, 157–177.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics, 28*, 267–283.

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance, 23*(2), 481–503.

Grossberg, S., Govindarajan, K. K., Wyse, L. L., & Cohen, M. A. (2004). ARTSTREAM: A neural network model of auditory scene analysis and source segregation. *Neural Networks, 17*(4), 511–536.

Grossberg, S., & Myers, C. W. (2000). The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review, 107*(4), 735–767.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language, 49*(1), 43–61.

Hannagan, T., Magnuson, J., & Grainger, J. (2012). A time-invariant connectionist model of spoken word recognition. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 1638–1643.

Hickok, G. S., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*, 393–402.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411–428.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology, 41*, 401–410.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition, 96*, B23–B32.

Kalénine, S., Mirman, D., Middleton, E. L., & Buxbaum, L. J. (in press). Temporal dynamics of activation of thematic and functional action knowledge during auditory comprehension of artifact words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics, 7*(3), 279–312.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*(2), 141–178.

Kraljic, T., Samuel, A. G ,. & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science, 19*, 332–38

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kukona, A., Fang, S., Aicher, K. A., Chen, H., & Magnuson, J. S. (2011). The time course of anticipatory constraint integration. *Cognition, 119*, 23–42.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*, 98–104.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior, 12*, 119–131.

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology, 55*, 306–353.

Lee, C-L., Middleton, E. L., Mirman, D., Kalénine, S., & Buxbaum, L. J. (in press). Incidental and context-responsive activation of structure- and function-based action features during object identification. *Journal of Experimental Psychology: Human Perception and Performance*.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

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

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54*(5), 358–368.

Lisker, L. (1985). The pursuit of invariance in speech signals. *Journal of the Acoustical Society of America, 77*, 1199–1202.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. (Research on Speech Perception, Technical Report No. 6). Bloomington: Speech Research Laboratory, Department of Psychology, Indiana University.

Luce, P. A., Goldinger, S. D., Auer, E. T., Jr., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and parsyn. *Perception and Psychophysics, 62*, 615–625.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*, 1–36.

Luce, R. D. (1959). *Individual choice behavior*. Oxford, England: Wiley.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676–703.

Magnuson, J. S. (2008). Nondeterminism, pleiotropy, and single word reading: Theoretical and practical concerns. In E. Grigorenko & A. Naples (Eds.), *Single word reading (pp. xx–xx)*. Mahweh, NJ: Erlbaum.

Magnuson, J. S., Dixon, J., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science, 31*, 133–156.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003a). Lexical effects on compensation for coarticulation: The ghost of Christmash past. *Cognitive Science, 27*, 285–298.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003b). Lexical effects on compensation for coarticulation: A tale of two systems? *Cognitive Science, 27*, 801–805.

Magnuson, J. S., Mirman, D., & Harris, H. D. (2012). Computational models of spoken word recognition. In M. Spivey, M. Joanisse, & K. McRae (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 76–103). Cambridge: Cambridge University Press.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 391–409.

Magnuson, J. S., Strauss, T., & Harris, H. D. (2005). Interaction in spoken word recognition models: Feedback helps. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 1379–1384).

Magnuson, J. S., Tanenhaus, M. K., & Aslin, R. N. (2008). Immediate effects of form-class constraints on spoken word recognition. *Cognition, 108*(3), 866–873.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The microstructure of spoken word recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General, 132*(2), 202–227.

Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*, 71–102.

Marslen-Wilson, W. (Ed.). (1989). *Lexical representation and process*. Cambridge, MA: MIT Press.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8*, 1–71.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words phonemes and features. *Psychological Review, 101*, 653–675.

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

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 576–585.

Matin, E., Shao, K., & Boff, K. (1993). Saccadic overhead: Information processing time with and without saccades. *Perception and Psychophysics, 53*, 372–380.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General, 134*, 477–500.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology, 23*, 1–44.

McClelland, J. L. (2010). Emergence in cognitive science. *Topics in Cognitive Science, 2*(4), 751–770.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences, 14*(8), 348–356.

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

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences, 10*(8), 363–369.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I An account of basic findings. *Psychological Review, 88*(5), 375–407.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86*, B33–B42.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language, 60*(1), 65–91.

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 433–443.

McQueen, J. M. (2003). The ghost of Christmas future: Didn't Scrooge learn to be good? Commentary on Magnuson,

McMurray, Tanenhaus, and Aslin (2003). *Cognitive Science, 27*(5), 795–799.

McQueen, J. M., Cutler, A., Briscoe, T., & Norris, D. (1995). Models of continuous speech recognition and the contents of the vocabulary. *Language and Cognitive Processes, 10*, 309–331.

McQueen, J. M., Jesse, A., & Norris, D. (2009). No lexical–prelexical feedback during speech perception or: Is it time to stop playing those Christmas tapes? *Journal of Memory and Language, 61*, 1–18.

McQueen, J. M., Norris, D., & Cutler, A. (2006). Are there really interactive processes in speech perception? *Trends in Cognitive Sciences, 10*(12), 533.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*, 227–234.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics, 25*, 457–465.

Mirman, D. (2008). Mechanisms of semantic ambiguity resolution: Insights from speech perception. *Research on Language and Computation, 6*(3–4), 293–309.

Mirman, D., Bolger, D. J., Khaitan, P., & McClelland, J. L. (in press). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cognitive Science.*

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory & Language, 59*(4), 475–494.

Mirman, D., & Graziano, K. M. (2011). Individual differences in the strength of taxonomic versus thematic relations. *Journal of Experimental Psychology: General.* doi:10.1037/a0026451

Mirman, D., & Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory and Cognition, 37*(7), 1026–1039. doi:10.3758/MC.37.7.1026

Mirman, D., McClelland, J. L., & Holt, L. L. (2005). Computational and behavioral investigations of lexically induced delays in phoneme recognition. *Journal of Memory and Language, 52*(3), 424–443.

Mirman, D., McClelland, J. L., & Holt, L. L. (2006). Reply to McQueen et al.: Theoretical and empirical arguments support interactive processing. *Trends in Cognitive Sciences, 10*(12), 534.

Movellan, J. R., & McClelland, J. L. (2001). The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review, 108*(1), 113–148.

Myung, J., Blumstein, S. E., & Sedivy, J. C. (2006). Playing on the typewriter and typing on the piano: Manipulation knowledge of objects. *Cognition, 98*, 223–243.

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

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*, 357–395.

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., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23*, 299–370.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*, 204–238.

Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology, 34*(3), 191–243.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 24, 175–184.

Peterson, G., & Lehiste, I. (1960). Durations of syllabic nuclei in English. *Journal of the Acoustical Society of America, 32*, 693–703.

Pisoni, D., & Tash, J. (1974) Reaction times to comparisons with and across phonetic categories. *Perception and Psychophysics, 15*(2), 285–290

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language, 39*, 347–370.

Pitt, M. A., & Samuel, A. G. (2006). Word length and lexical activation: Longer is better. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 1120–1135.

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney (Ed.), *The emergence of language* (pp. 381–415). Mahwah, NJ: Erlbaum.

Pustejovsky, J. (1995). *The generative lexicon.* Cambridge, MA: MIT Press.

Rogers, T. T., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. E. (2004). Natural selection: The impact of semantic impairment on lexical and object decision. *Cognitive Neuropsychology, 21*(2–4), 331–352.

Rubin, P., Turvey, M. T., & van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception and Psychophysics, 19*, 384–398.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90*, 51–89.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110*, 474–494.

Samuel, A. G. (1996). Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General, 125*(1), 28–51.

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology, 32*(2), 97–127.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science, 12*, 348–351.

Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language, 48*(2), 416–434.

Saslow, M. G. (1967). Latency for saccadic eye movement. *Journal of the Optical Society of America, 57*, 1030–1033.

Shillcock, R. C., & Bard, E. G. (1993). Modularity and the processing of closed class words. In G. T. M. Altmann & R. C. Shillcock (Eds.), *Cognitive models of speech processing. The Second Sperlonga Meeting* (pp. 163–185). Mahwah, NJ: Erlbaum.

Spivey, M., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences USA, 102*(29), 10393–10398.

Strauss, T., Harris, H. D., & Magnuson, J. S. (2007). jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behavior Research Methods, Instruments and Computers, 39*, 19–30.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of con- text effects. *Journal of Verbal Learning and Verbal Behavior, 18*, 645–659.

Tanenhaus, M. K., Leiman, J. M., & Seidenberg, M. S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior, 18*, 427–440.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 632–634.

Trueswell, J. C., & Tanenhaus, M. K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives in sentence processing* (pp. 155–179). Hillsdale, NJ: Erlbaum.

Tyler, L. K., Voice, J. K., & Moss, H. E. (2000). The interaction of meaning and sound in spoken word recognition. *Psychonomic Bulletin and Review, 7*, 320–326.

Ueno, T., Saito, S., Rogers, T. T., & Lambon Ralph, M. A. (2011). Lichtheim 2: Synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron, 72*(2), 385–396.

Vitevitch, M. S., & Luce, P.A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40*, 374–408.

Viviani, P. (1990). Eye movements in visual search: Cognitive, perceptual, and motor control aspects. In E. Kowler (Ed.), *Eye movements and their role in visual and cognitive processes. Reviews of oculomotor research (Vol. 4, pp. xx–xx).* Amsterdam, The Netherlands: Elsevier.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*, 392–93.

Warren, R. M., Obusek, C. J., Farmer, R. M., & Warren, R. P. (1969). Auditory sequence: Confusion of patterns other than speech and music. *Science, 196*, 586–587.

Yee, E., & Sedivy, J. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*(1), 1–14.

Yee, E., Huffstetler, S., & Thompson-Schill, S. L. (2011). Function follows form: Activation of shape and function features during object identification. *Journal of Experimental Psychology: General, 140*(3), 348–363.

Yee, E., Overton, E., & Thompson-Schill, S. L. (2009). Looking for meaning: Eye movements are sensitive to overlapping semantic features, not association. *Psychonomic Bulletin and Review, 16*(5), 869–874.

Zhuang, J., Randall, B., Stamatakis, E. A., Marslen-Wilson, W. D., & Tyler, L. K. (2011). The interaction of lexical semantics and cohort competition in spoken word recognition: An fMRI study. *Journal of Cognitive Neuroscience, 23*(12), 3778–3790.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition, 32*, 25–64.