Chapter 8
Spoken Word Recognition

Delphine Dahan and James S. Magnuson

1. INTRODUCTION

We solve an astounding array of information-processing challenges when we perceive a speaker's intended message. Apparently effortlessly, we accommodate variability in talker characteristics, dialect, speaking rate, and acoustic environment, all of which perturb the mapping between speech and linguistic categories. Without the aid of invariant cues to phonetic categories or word boundaries, we map acoustics onto phonetic categories, phonetic categories onto words in memory, words onto phrases and syntactic structures, words and syntax onto semantics, etc. Or do we?

On this view of language understanding, spoken word recognition is a distinct subsystem providing the interface between low-level perception and cognitive processes of retrieval, parsing, and interpretation. The narrowest conception of the process of recognizing a spoken word is that it starts from a string of phonemes, establishes how these phonemes should be grouped to form words, and passes these words onto the next level of processing. Some theories, though, take a broader view and blur the distinctions between speech perception, spoken word recognition, and sentence processing (e.g., Elman, 2004; Gaskell & Marslen-Wilson, 1997; Klatt, 1979; McClelland, St John, & Taraban, 1989).

What motivates the narrow and broad conceptions? There are empirical, pragmatic, and theoretical motivations for the narrow view. Empirically, psycholinguistic levels of processing map roughly onto linguistic levels of description. The fact that linguistic knowledge can be described as a hierarchically structured set of levels leads to the reasonable hypothesis that speakers (or signers) and perceivers may represent and operate on those structures. Indeed, this hypothesis is given face validity by the fact that humans can make decisions about levels like phonemes and words and that perception can be influenced by manipulations at those levels (though there is a long history of debate over their psychological reality; see Pisoni & Luce, 1987, for a review).

Handbook of Psycholinguistics
ISBN: 0-12-369374-8

The pragmatic motivation for the narrow view stems from the fact that over a century of concerted study of speech perception has led to a catalog of complex empirical phenomena and candidate cues for speech perception, but little understanding of the specific components of the speech signal that humans use to decode speech and achieve phonetic constancy (Nusbaum & Magnuson, 1997; Remez, 2005). Rather than wait for a complete understanding of early perceptual processes, psycholinguists have made significant progress in understanding the processing of words and sentences by making the simplifying assumption that a string of phonemes makes a reasonable proxy for the results of initial perception, and that a series of sound forms associated with lexical entries makes a reasonable proxy for the input to sentence processing.

Theoretically, the narrow view is motivated in part by the assumption that the division of labor in staged systems affords significant processing efficiencies (Fodor, 1983; Norris, 1994; Norris, McQueen, & Cutler, 2000). Breaking the problem into distinct stages is argued to provide cognitive economy if the result is a series of mappings that are straightforward relative to the complexity of the full mapping from lowest to highest level (restrictions on the information available to each level are also key to the *interaction vs. autonomy* debate discussed below).

The broader view of spoken word recognition (in the extreme, as the mapping from speech to meaningful units that may be larger than words) has empirical and theoretical motivations. One consideration is that by assuming that the input to spoken word recognition is a string of abstract, phonemic category labels, one implicitly assumes that the nonphonemic variability carried on the speech signal is not relevant for spoken word recognition and higher levels of processing. However, if this variability and detail is not random but is lawfully related (even partially) to linguistic categories, the simplifying assumption that the output of speech perception is a string of phonemes may actually be a complicating assumption. Indeed, there is growing evidence that spoken word recognition is influenced by information in the signal that cannot be captured in a string of phonemes. For example, misleading coarticulatory cues caused by splicing the onset and most of the vowel of one consonant–vowel–consonant (CVC) word or nonword onto the last consonant of another CVC word (creating “subcategorical mismatches”; Whalen, 1984) changes the time course of lexical activation and competition (Dahan, Magnuson, Tanenhaus, & Hogan, 2001a; Marslen-Wilson & Warren, 1994; McQueen, Norris, & Cutler, 1999).

What purpose might this fine-grained sensitivity serve? One challenge posed by assuming that words are identified from a string of phonemes is the *embedding problem*; most long words have multiple shorter words embedded within their phonemic transcriptions (e.g., depending on dialect, and neglecting all subphonemic cues, *unitary* contains *you, unit, knit, it, tarry, air,* and *airy*) and conversely, many short words embed in one or more other words (McQueen, Cutler, Briscoe, & Norris, 1995). Successful spoken word recognition depends on distinguishing intended words from embeddings. However, the embedding problem is significantly mitigated when subphonemic information in the input is considered. For example, listeners are sensitive to very subtle durational differences (in the range of 15–20 ms) that distinguish *phonemically identical* syllables that occur in short words (*ham*) from those embedded in longer words (*hamster*) (Salverda, Dahan, & McQueen, 2003; see also Davis, Marslen-Wilson, & Gaskell, 2002).

Thus, the bottom-up signal contains vital information that simplifies the mapping from speech to words that would be lost were words identified from a string of phonemes. Might the same be true for subsequent processes? There is increasing evidence that the construction of syntactic and semantic structures relies on more than just a sequence of words. Indeed, a sequence of words is almost always temporarily compatible with multiple structures. For example, the structure associated with the word sequence *John knew the answer* differs whether it is followed by *was wrong* or *to the question* (e.g., Altmann, 1999). A growing body of work has documented the role played by the prosodic structure of an utterance (marked by prosodic breaks and intonational prominences) in favoring some structures over others (e.g., Kjelgaard & Speer, 1999; for a review see Cutler, Dahan, & van Donselaar, 1997; Speer & Blodgett, 2006, this volume). This literature indicates that information from the speech signal is passed onto higher levels of processing. This supports an integrated view of phonetic, lexical, and sentential processing.

Sentence-level top-down constraints on lexical activation have received some attention in spoken word recognition, but chiefly with respect to how top-down information might constrain the set of activated lexical items (e.g., Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987, 1990). Immediate access to syntactic, semantic, and nonlinguistic context could provide significant constraints on spoken word recognition, by influencing the activation of homophones, semantic associates, or context-appropriate lexical items (Shillcock & Bard, 1993), helping resolve lexical ambiguity resulting from phonological assimilations (Gaskell & Marslen-Wilson, 2001), or by restricting the set of possible referents (Brown-Schmidt, Campana, & Tanenhaus, 2004).

Throughout this chapter, as we describe the central themes of current research on the recognition of spoken words, we will adopt the more prevalent, narrow view (except when noted) that most current work assumes. As we will discuss at the end of the chapter, the growing evidence for subcategorical specificity may herald a dramatic shift in theories of spoken word recognition. Taking the broad view – confronting the speech signal itself and considering how higher levels of representation might constrain lexical access - may be the key to significant progress in understanding spoken word recognition.

The recognition of a spoken word can be viewed as the process of classifying an auditory stimulus as belonging to one “word-form” category, chosen from many alternatives. As this description stresses, this process requires matching the spoken input with mental representations associated with word candidates, and selecting one among several candidates that are at least partially consistent with the input. Frauenfelder and Tyler (1987) classified the functions required of any theory of spoken word recognition into three stages. *Initial contact* is how input interfaces with and activates lexical representations. *Selection* describes how the set of activated lexical alternatives is evaluated with respect to the sensory input. *Integration* refers to how candidates are evaluated with respect to the linguistic and nonlinguistic context, in order to identify which is the likeliest candidate for recognition as well as to build larger linguistic structures.

Early models viewed these processes as discrete, only partially overlapping stages, in particular predicting a temporal delay between access and selection (e.g., Marslen-Wilson, 1987; Zwitserlood, 1989). More recent models allow for continuous uptake and evaluation of the input, thus blurring functional and temporal distinctions between access, selection, and integration (for behavioral evidence supporting the continuous view, see Dahan & Tanenhaus, 2004). Nonetheless, the theoretical distinctions are useful as each process poses challenges of a different nature.

In the course of this chapter, we will review the central issues pertaining to contact and selection in turn, and how they have been conceptualized in different models and theories of spoken word recognition. We will also address whether categorizing the input as a member of a word category changes listeners' percept of the input. This question hinges on the architecture of the processing system, i.e., whether higher levels of representations (such as words) can affect lower levels, such as speech sounds or phonemes. Finally, we will briefly review integration, and close with a discussion of what we see as the most crucial challenges to theories of spoken word recognition and spoken language generally, and the approaches we find most promising and most likely to lead to solutions.

2. INITIAL CONTACT

When someone speaks, the linguistic content and speaker characteristics (e.g., physiology of the vocal tract, gender, regional origin, emotions, identity) simultaneously influence the acoustics of the resulting spoken output. Additional sources of variability include rate of elocution, prosodic prominence, and the phonetic context in which each word is pronounced. Nonetheless, listeners are able to recognize acoustically different stimuli as instances of the same word, thus extracting the similarity that exists between these different tokens, and perceiving them as members of the same category. How are words mentally represented to allow for this complex categorization?

The traditional (and dominant) view assumes that people represent the form of words as categories that abstract away from variability. Drawing on linguistic theories, the mental representation of a word form is usually conceived as a sequence of phonemes (sometimes themselves decomposed into a bundle of contrastive features). Within this framework, the ease with which a given pronunciation is categorized as a token of a given word is assumed to depend upon the degree to which its components have characteristics typically associated with the word's phonemes. Speaker-specific information is often viewed as a source of noise which does not contribute to the process of identifying the linguistic units present in the signal.

This view has not gone uncontested. An episodic view, most forcefully argued for by Goldinger (1996, 1998), conceptualizes lexical representations as ensembles of detailed memory traces (or episodes) of word instances. Several recognition memory studies have shown that people implicitly retain in memory nonlinguistic aspects of spoken words (e.g., Hintzman, Block, & Inskeep, 1972; Goldinger, 1996, 1998; Palmeri, Goldinger, & AQ1 Pisoni, 1993). The question at stake is whether these memory traces of words constitute the knowledge that people access and use when identifying spoken words. Goldinger (1998) applied (Hintzman's) (1986) MINERVA2 model of episodic memory to spoken AQ2 word recognition. In this model, a speech episode (a word) is simultaneously compared to all memory traces. Activation of a trace is proportional to its acoustic similarity with the stimulus. The aggregate of all activated traces (the so-called *echo*) is sent to working memory and corresponds to the listener's percept. Because the echo consists of a blend of the memory traces that resemble the stimulus, it tends to capture the aspects that are common among the traces but not the aspects that differ. This principle enables the model to make generalizations and categorize new tokens without assuming the existence of abstract mental categories. A critical challenge to the episodic view is how the similarity between an actual speech stimulus and memory traces would be computed, if no normalization or other data-reducing process abstracting from surface variability is assumed. Goldinger's model has thus far assumed word-length episodes and remains agnostic about how words would be isolated from the utterances they are embedded in, which is problematic given the challenges posed by word segmentation (see below). Given its radical departure from classical approaches, this theory may well have the potential to bring new leverage to problems of speech perception and spoken word recognition. However, until similarity mapping and segmentation are spelled out, the episodic view faces the same challenges as the traditional, abstract view.[1]

The traditional view has influenced much of the research on spoken word recognition. Thus, the recognition of a spoken word is generally viewed as the mapping of the speech input onto abstract lexical representations, with abstract units standing for the word's subcomponents, the phonemes, mediating this mapping. An extended line of research has documented how listeners accommodate the variability inherent to speech rate and phonetic context in the perception and recognition of individual phonemes (Miller & Liberman, 1979). We will not review this literature here, but rather will focus on how theories of spoken word recognition have embodied, or sometimes departed from, the classical approach to spoken word recognition.

[1] Goldinger's (1998) simulations have two critical problems. The model assumes the input is pre-segmented into word-length episodes (the primitive unit), which are represented as vectors of units (with values of −1, 0, or +1), with some units representing the word type, and others representing talker and context information. While Goldinger claimed such a model can achieve phonetic constancy without normalization of talker differences, the solution depends on this unrealistic assumption about the input. In real speech, talker variability conditions phonetic realization. In Goldinger's simulations, the input is in effect pre-normalized. The episodic model's promise to solve phonetic constancy without normalization may be possible, but tests with more realistic input are needed to evaluate it (see Goldinger & Azuma, 2003, for a discussion of how adaptive resonance theory may provide the means to test the theory with more realistic inputs and mechanisms).

2.1. Initial Contact and Similarity Metrics

The first question for any model is the nature of the input representation: How do the products of sensory information interface with the lexicon? As mentioned earlier, the input to word recognition has traditionally been assumed to be a string of phonemes, output by a speech perception system (as in the original COHORT model; Marslen-Wilson & Welsh, 1978). This representation was also adopted by the model SHORTLIST (Norris, 1994), although mainly for practical reasons. However, the string-of-phonemes encoding of the speech input assumes that subphonemic variation in the signal is lost, while such variation has been shown to affect listeners' word recognition. For example, Andruski, Blumstein, and Burton (1994) demonstrated that, as the realization of the initial segment of a word like *king* is modified as to differ from a prototypical /k/ (by shortening the duration of the stop voice onset time) but not enough to change the ultimate categorization of this segment, people are nonetheless less likely to categorize the word as an instance of *king* (see McMurray, Tanenhaus, & Aslin, 2002, for converging evidence; for demonstrations of listeners' sensitivity to subcategorical cues in vowels, see (Dahan et al., 2001a,b); Marslen-Wilson & Warren, 1994; Whalen, 1984). Evidence for graded activation of words based on subphonemic similarity requires a finer representational grain than phonemes.

Another issue related to the string-of-phonemes assumption is that it imposes a dissociation between the process of recognizing words from that of recognizing its components. A recent attempt to add an automatic phone recognizer to the SHORTLIST model exposed the limitations of this assumption (Scharenborg, ten Bosch, Boves, & Norris, 2003). In these simulations, the automatic phone recognizer took real speech (naturalistic speech samples from telephone conversations) as input and generated a sequence of phonemes. From this string of phonemes, the activation and competition mechanisms implemented in the SHORTLIST model yielded the best matching word candidate. Word recognition accuracy performance was poor (around 25%), but improved considerably when one of the model's parameters, the penalty assigned to candidates that mismatch the phonemic input, was set to zero. This result may be interpreted as evidence that SHORTLIST, originally tested on an unrealistically accurate phonemic input, must be revised to accommodate likely erroneous input from a phone recognizer. On the other hand, this result can be taken as reflecting the shortcoming of a phonemic string as input. If "hard" phonemic decisions are made by the phone recognizer, the fact that other phonemic interpretations were substantially supported by the signal is lost. Most of all, these simulations illustrate how much the modeling of spoken word recognition hinges on assumptions about the input representation.

The next simplest solution is to assume that the input takes the form of localist phoneme activation units (as in the REVISED COHORT model (Marslen-Wilson, 1987, 1989), and MERGE (Norris et al., 2000)). Subphonemic detail can be approximated in the distributed representation afforded by the entire set of phonemes. Thus, a segment ambiguous between /k/ and /g/ can be represented by partial activation of both units. A slightly more fine-grained representation can be achieved with units representing a set of (usually binary) acoustic-phonetic features (as in the DISTRIBUTED COHORT MODEL

(Gaskell & Marslen-Wilson, 1997)). However, the realism of these schemes is limited, as they fail to incorporate a critical aspect of speech, coarticulation.

As demonstrated by the seminal work by Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967) and contrary to a listener's subjective impression, a spoken utterance is not a concatenated sequence of discrete speech sounds. The gestures involved in the articulation of each sound overlap temporally with the gestures that generate adjacent sounds. One of the consequences of this temporal overlap has been coined the "segmentation" problem. A spoken utterance cannot be divided into smaller portions, each one representing a single segment.[2] If the recognition of a spoken word involves the mapping of the input onto word representations where segments are in temporal order, the listener must assign the presence of a given acoustic feature in the input to a given segment.

The TRACE model (McClelland & Elman, 1986) uses the most complex input representations of any current model of speech perception and spoken word recognition. The input is a "pseudo-spectral" representation based on seven acoustic-phonetic features, each represented with a nine-unit continuous vector, which encode the degree to which the feature is represented in the input. Features spread over time by ramping up to a phoneme center and then ramping off. Phoneme centers are close enough together and features spread far enough that there is substantial overlap between phonemes, creating a rough analog to coarticulation.

TRACE's architecture is also a critical aspect in the way it accounts for the processing of coarticuated speech. In TRACE, units that stand for hypotheses at the featural, phonemic, or word level, are replicated every three time slices. Each unit stands for a linguistic unit potentially present in the input at a different point in time. The extensive unit reduplication has often been criticized as an implausible feature of the model (beginning with McClelland & Elman, 1986). However, this is central to solving the segmentation issue, as it accommodates the fact that features that result from the overlap of articulatory gestures coincide in time. A given time slice may provide evidence supporting different phonemes, thus activating several incompatible phoneme units. However, within each level, units that span the same portion of the input inhibit each other. Consequently, the phoneme hypothesis for which the evidence is the strongest can win the competition. Thus, TRACE 's architecture allows the segmentation of coarticulated speech into a sequence of discrete segments.

However, TRACE's input scheme provides a very rough approximation of coarticulation in real speech. While it accommodates the temporal overlap of gestures, it fails to accommodate the fact that this temporal overlap affects the articulatory (and therefore acoustic) realization of segments (i.e., the "lack of invariance" issue, cf. Liberman et al., 1967). There have been very few even moderately successful attempts to devise

[2] This is true despite the fact that the duration of single segments, such as consonants or vowels, are often reported. Such segmentation is based on conventions on how to define boundaries between segments based on their relative prominence (see Fowler, 1984).

psychologically tractable models that work directly on the actual speech signal or a minimally transformed speech signal. (The hidden-Markov models and similar mechanisms used in automatic speech recognition systems arguably substitute the black box of the brain with largely opaque statistical approximations; see Nusbaum & Magnuson, 1997, for discussion.) Klatt's (1979) LEXICAL ACCESS FROM SPECTRA (LAFS) model is perhaps the best known, but the mapping from spectra to lexical items is at least as variable as the mapping from speech to phonemes. Work in the adaptive resonance framework has grappled with real speech signals (the ARTSTREAM model; Grossberg, Govindarajan, Wyse, & Cohen, 2004) but has yet to be extended to the recognition of phonemic or lexical forms. The strategy of Plaut and Kello (1999) may well be the best hope for progress toward more realistic input. They use a collection of articulatory and acoustic cues that *might* turn out to be tractable to extract from speech (auditory and visual cues to jaw movements, changes in formants, etc.), and in combination, *might* prove a sufficient basis for speech perception and spoken word recognition.

2.2. Initial Constraints on Activation

Theories differ on the patterns of activation that follow initial contact. More specifically, they differ in the theories of similarity they assume. The ORIGINAL (Marslen-Wilson & Welsh, 1978), REVISED (Marslen-Wilson, 1987, 1989) and DISTRIBUTED COHORT models (Gaskell & Marslen-Wilson, 1997, 1999, 2002) place great emphasis on word onsets. The real-time constraints of the speech signal motivate an emphasis on optimal use of bottom-up information as it becomes available. Since a word's onset is heard first, it should determine which lexical items are first activated. Thus, in the original COHORT model, the set of activated lexical alternatives was constrained to a *word-initial cohort* of items that matched perfectly the phonemic representation of the first approximately 150 ms of a word's onset. In light of evidence that a word might be recognized even when its first sounds are altered (for example, due to mispronunciation, cf. Cole, 1973), the revised and DISTRIBUTED COHORT models abandon the strict, all-or-none match constraint. Instead, lexical representations are activated as a function of their similarity to a spoken word, with this similarity being continuously evaluated rather than limited to the initial portion of the spoken word. Nonetheless, the models' emphasis on real-time processing maintains a special status to the spoken word's initial sounds, as they contribute to the activation of some words, and thereby the interpretation of subsequent spoken material will be biased in favor of these words (see the discussion of *Selection* below for a full description of how these biases might be implemented).

The NEIGHBORHOOD ACTIVATION MODEL (NAM; Luce, 1986; Luce, Pisoni, & Goldinger, 1990; Luce & Pisoni, 1998) differs from any instantiation of the COHORT model by predicting activation of words that reflects their global similarity with the spoken word.[3]

[3] NAM is not a processing model *per se* – it is more properly considered a formal similarity model. However, its similarity metric imposes significant constraints on an underlying processing mechanism, and as such, it is appropriate to consider what NAM predicts in terms of lexical activation.

Two similarity metrics were developed within the model. The more complex one is derived from observed similarity measures, such as position-specific diphone confusion probabilities. Similarity between the spoken word and other words is computed by comparing the confusability of each of its segments with other words' segments in the same position within the word. Similarity is gradient, although limited to words that have the same number of segments. The simpler metric, sometimes called the *one-phoneme shortcut metric*, distinguishes words that are predicted to become activated during the perception of a spoken word (i.e., its neighbors) from those that are not, with no gradiency in the degree of activation of the former. Activated words (i.e., neighbors of the spoken words) are defined as words that differ from the spoken word by no more than one phoneme, whether by substitution, deletion, or addition, in any position. Thus, neighbors of *cat* include *bat*, *kit* and *cap* (substitutions), *at* (deletion), and *scat* and *cast* (additions).

The COHORT and NEIGHBORHOOD models make different predictions about what items may be activated by a spoken word. The COHORT model predicts that hearing *cat* also activates *castle* but should activate *bat* to a negligible degree. NAM predicts that *cat* will activate *bat* but not *castle*, as it differs by too many phonemes. There is empirical support for each prediction. Marslen-Wilson (1993) reported a series of studies in which the auditory presentation of a word primes visual lexical decisions to semantic associates of words overlapping in onset, but not in rhyme (e.g., *beaker* would prime *insect*, an associate of *beetle*, but not *stereo*, an associate of *speaker*). But Luce and Pisoni (1998) reported that neighborhoods based on global similarity provide the best prediction of processing time for large sets of words in tasks like lexical decision and naming, although they did not separate out the contribution of the cohort-type neighbors from that of non-cohort ones (we discuss this result further in the *selection* section).

TRACE makes an intermediate prediction: It activates both onset- and rhyme-overlapping words, because, as in the NEIGHBORHOOD model, words can be activated even if they mismatch at onset. However, unlike the NEIGHBORHOOD model, TRACE represents time: Words that become activated early in the spoken input have an advantage over words that become activated later, because more of the spoken word has been heard and selection mechanisms are then more effective at favoring the best matching candidate. Thus, TRACE predicts activation of both onset- and rhyme-overlapping candidates, although at different times and of different amplitude. Allopenna, Magnuson, and Tanenhaus (1998) provided behavioral data supporting this prediction. They estimated lexical activation to word candidates by monitoring eye movements to pictures as participants followed verbal instructions to move an item on a computer screen. Fixations were closely time-locked to the speech (with a lag only slightly larger than that required to plan and launch an eye movement), and mapped closely onto phonetic similarity over time (with higher and earlier fixation proportions to onset-overlapping competitor than rhyme-overlapping competitor) as well as response probabilities generated by TRACE. This study highlights the importance of a measure of lexical activation over time, given the rapid evolution of lexical activation over time, as the spoken input is heard.

The Allopenna et al. (1998) study highlights one shortcoming of the similarity model embodied in NAM to the study of spoken word recognition. The temporal distribution of similarity is not considered; *dab* and *bad* are assumed to be equally active upon hearing *dad* (ignoring frequency for the sake of the example). NAM fails to capture the temporal dimension of speech and the special status that the initial sounds have due to their temporal precedence (Marslen-Wilson & Zwitserlood, 1989). It also gives too much weight to the match in the number of segments or syllabic structure by entirely excluding the contribution of words that are more than one phoneme longer than the word to be recognized, despite evidence suggesting that words of different lengths affect the processing of a given word (Marslen-Wilson, 1984, 1987). The algorithm cannot be easily extended to the computation of competition environment for polysyllabic words, as most of these words have very few, if any, competitors under the one-phoneme difference definition.[4] Finally, the one-phoneme shortcut metric, which has been most widely used by researchers and has proven useful in stimulus selection and experimental control, treats any phoneme deviation equally, regardless of its phonetic nature. Confusion between two words differing by one-phoneme addition or substitution, or confusion between two words differing by a vowel or a consonant, are all assumed to be equivalent, despite empirical evidence that the nature of the phonetic feature(s) that differ between two words is an important factor in accounting for word confusions (e.g., Bailey & Hahn, 2005; Hahn & Bailey, 2004; see also van Ooijen, 1996).[5]

2.3. Plasticity in Mapping the Speech Signal onto the Lexicon_

As pointed out in the introduction to this section, the acoustic form that a given word takes can vary greatly. Nonetheless, listeners have little difficulty accommodating this variability, which has sometimes been interpreted as reflecting plasticity in the mapping of the speech signal onto the lexicon. Here we review some of this work.

A substantial number of studies have examined the processing of spoken words that have undergone phonological assimilation. In connected speech, the value of a segment's feature (i.e., place of articulation or voicing) may assimilate to that of the same feature from its surrounding segments. For instance, the place of articulation of the final alveolar sound of the word *green*, may be altered to become (or approach) the bilabial place of articulation of the initial sound of the subsequent word *boat*, so that the sequence may sound a little like *gream boat* The conditions under which assimilation may occur are dictated by the phonology of the language. Research on the perception of assimilated words

[4] Cluff and Luce (1990) used the one-phoneme difference algorithm to compute the competition environment of bisyllabic words composed of two monosyllabic words (e.g., jigsaw) by establishing competitors for each syllable independently, thereby considering only monosyllabic competitors.

[5] Luce, Goldinger, Auer, and Vitevitch (2000) report examples of cases where a more complex metric, based on positional similarity ratings, makes distinctly different predictions than the one-phoneme shortcut metric, e.g., predicting competition between *veer* and *bull* due to high similarity at every segment despite no complete phoneme matches.

has shown that this deviation does not preclude the identification of the assimilated token as an instance of the intended word. Gaskell and colleagues (Gaskell, 2003; Gaskell, Hare, & Marslen-Wilson, 1995; Gaskell & Marslen-Wilson, 1996, 1998) have suggested that listeners have learned to accept the assimilated form as a token of the intended word in the appropriate context, especially if the assimilation was only partial (thus, maintaining some of the acoustic characteristics of the original segment). This proposal was supported by simulations from a connectionist model that was trained to learn to map acoustically variable (but arguably simplified) input onto canonical, fixed representation of words (see Gow, 2001, 2002, 2003a, 2003b, for a critique of Gaskell and colleagues' proposal and for a competing account of the perception of assimilated words).

Other rule-based variations do not involve a phonemic or subphonemic alteration. These pronunciations are characteristic of casual (as opposed to careful) speech, and often described as including an atypical or reduced realization of some of the segments of words. For example, the final consonant of the word *flute* can be realized with an alveolar closure and an audible release (the typical realization of the phoneme /t/), or realized as a glottal stop, with no release. Similarly, the vowel of the unstressed syllable of a polysyllabic word can be so drastically reduced that it is not acoustically present in the signal (e.g., *police*, pronounced roughly as [plis]). How can two fairly different realizations be interpreted as instances of the same word? Do people represent the multiple forms that a word can take in order to accommodate such variation? And if so, at what level of abstraction are they represented, and does frequency of occurrence of variants determine whether a variant is represented or not?

Recent work has addressed these questions by examining whether variations are equally effective at making contact with the lexical representation of the intended word (or, put slightly differently, whether variations are equally categorized as members of the intended word category). Some studies have probed the degree to which the meaning of the word becomes available (e.g., Deelman & Connine, 2001; Sumner & Samuel, 2005). Other studies have examined the degree to which variants map onto the same or different representations by assessing whether having heard one variant facilitates the subsequent processing of the alternative (e.g., LoCasto & Connine, 2002; McLennan, Luce, & Charles-Luce, 2003; Sumner & Samuel, 2005; Utman, Blumstein, & Burton, 2000). The findings that emerge from these studies are complex and often conflicting. Some results suggest that any variant facilitates the processing of any alternative, which is sometimes interpreted as evidence for a single, abstract and general representation; other results argue for specificity. Some researchers found evidence for a special status of the most frequent variant (Connine, 2004), while others did not (Sumner & Samuel, 2005). This line of research has only begun, and it is too early to draw definite conclusions. Nonetheless, one aspect that has been little considered is the relevance of the context in which these variants occur. Listeners may be sensitive to how likely and expected a given variation is, given what is known of the talker's speaking style, speed of elocution, and perhaps geographic or dialectal origin. Such expectations (or use of context) may determine the degree to which a token will be mapped onto or activate representation(s) associated with the intended word.

Indeed, we know that listeners do adapt to the characteristics of the talker or speech that they hear. Evidence of such adaptation comes from studies showing that word identification is impaired by trial-to-trial changes in the voice of the talker (Mullennix, Pisoni, & Martin, 1989; Nusbaum & Morin, 1992) and/or or in his/her speaking rate (Sommers, Nygaard, & Pisoni, 1994), and from studies showing advantage for the identification of words spoken in a familiar vs. unfamiliar voice ((Nygaard), Sommers, & Pisoni, 1994; although see Luce & Lyons, 1998). This suggests plasticity in the process of perceiving and interpreting speech. Listeners' ability to adapt to the characteristics of the speech or the talker they are exposed to has long been acknowledged (e.g., Joos, 1948; Ladefoged & Broadbent, 1957; Peterson & Barney, 1952). More recently, a number of studies have documented how adaptation to distorted or foreign-accented speech proceeds. The process appears to operate quite rapidly, with measurable improvement in comprehension observed after as little as two to four sentences (Clarke & Garrett, 2004). Importantly, from relatively short exposure to distorted speech, people acquire knowledge that can generalize to sentences containing unheard words (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Greenspan, Nusbaum, & Pisoni, 1988), or to similarly distorted speech from a different talker (e.g., Dupoux & Green, 1997). Furthermore, listeners' perceptual adaptation to unusual speech or talker characteristics seems to be (at least largely) mediated by lexical knowledge. Listeners who were exposed to 20 distorted nonsense sentences prior to testing on sensible sentences fared no better than people with no prior exposure to distorted speech (Davis et al., 2005; for similar conclusions, see Eisner & McQueen, 2005; Norris, McQueen, & Cutler, 2003).

AQ4

Evidence for plasticity in the mapping of spoken input onto lexical representations may help explain how listeners cope with the extreme variability found in speech. So long as this variability is context-dependent, and thus lawful, prior (even brief) exposure to speech from a new talker may trigger the learning of a new mapping between speech input and linguistic units.

2.4. The Interaction Debate: Is the Interface Bidirectional?

Is the architecture underlying spoken word recognition autonomous (feedforward only) or interactive (lexical representations feed information back over the interface to the sublexical representations)? Bottom-up and top-down information *is* integrated: The literature is full of examples of lexical effects on tasks that tap sublexical representations. Phonemes are detected more quickly in words than nonwords (the word superiority effect; Rubin, Turvey, & Van Gelder, 1976). Listeners report hearing phonemes consistent with lexical or sentential context in locations completely replaced with noise (the phoneme restoration effect; e.g., Warren, 1970; Samuel, 1981, 1997). If a phoneme continuum is attached to a context that makes one endpoint a word and the other a nonword (e.g., /t/-/d/ attached to *–ash* or *–ask*), categorical perception boundaries shift such that more steps are identified as consistent with the lexical endpoint (Ganong, 1980; a bias is also found in word–word contexts with a frequency differential; Fox, 1984). Helpful visual contexts are integrated quickly to resolve ambiguities in sentence processing

(Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). The crux of the interaction debate is *when* integration occurs.

This debate has recently taken center stage in spoken word recognition research, having been energized by forceful empirical and theoretical arguments for autonomous models of spoken word recognition by Norris et al. (2000, 2003). In brief, the *autonomous view* is that processing stages can be optimized by allowing them access only to bottom-up information (disallowing interaction of top-down information). This view of stages within a processing system is related to arguments for modularity between processing systems (Fodor, 1983). In both cases, the idea is that veridical perception depends upon transparent processing of the incoming signal. On this view, if top-down information is integrated directly with sensory information, an organism *ipso facto* loses the possibility of veridical perception, as there is no distinction between information in the environment and information in the organism. Autonomous models account for lexical effects on sublexical tasks by proposing parallel, competing lexical and sublexical routes (as in the Race model; Cutler & Norris, 1979), or that the locus of sublexical decisions is, counter-intuitively, post-lexical. In the Merge model (Norris et al., 2000), for example, there are two banks of phoneme units. One is the source of bottom-up input to the lexical layer. The second receives input from the bottom-up phoneme nodes and the lexical nodes. This decision layer can thus integrate lexical and phonological knowledge without changing the prelexical interpretation of the sensory input. The separate bank of decision nodes is justified on the grounds that phonemic awareness is a late-developing artifact of learning to read, based on evidence that phonemic awareness does not develop if one does not learn to read (see Norris et al., 2000; Marslen-Wilson & Warren, 1994, for discussion; but there is evidence that sublexical awareness (if not precisely phonemic) does emerge in preliterate children (see Liberman, Shankweiler, Fischer, & Carter, 1974) and illiterate adults (Bertelson & de Gelder, 1989); see Shankweiler & Fowler, 2004, for a review). Falsifying this position would require showing that top-down lexical effects have a perceptual, rather than decisional, locus.

On the *interactive view*, if top-down information *can* usefully constrain interpretation of bottom-up information, it should be used, and veridical perception can be maintained by properly weighting bottom-up and top-down information. Falsifying this position is more difficult. Alternative explanations for lexical effects must be proposed, and evidence must show that when those explanations make predictions that are different from lexical feedback predictions, the lexical feedback predictions are incorrect. Over the past two decades, the debate appeared to be settled at least two or three times, with alternative apparent falsifications of autonomous and interactive positions.

Elman and McClelland (1988) seemingly falsified the autonomous position, by showing lexical effects on sublexical processing rather than sublexical decisions. They conducted a study designed to demonstrate *lexically mediated compensation for coarticulation*. Compensation for coarticulation (Mann & Repp, 1981) refers to the fact that in normal production, if a segment with a front place of articulation follows one further back (or vice versa), physical and temporal constraints may prevent the articulation from reaching its ideal location, with the result that in this context, the front segment will have a place of articulation further back than normal. When a front-back continuum (e.g., /t/-/k/) is presented following a back segment (e.g., /Σ/) the category boundary shifts toward the back (i.e., more steps on the continuum are identified as the +front segment /t/), and the opposite happens after a front segment (e.g., /s/). In Elman and McClelland's (1988) study, this low-level perceptual phenomenon was coupled with the Ganong (1980) effect. The Ganong effect shows that the interpretation of an ambiguous sound (symbolized by ?, intermediate between, e.g., *p* and *b*) embedded in a larger spoken stimulus (e.g., ?*eace*) is biased toward the interpretation that turns the spoken stimulus into a real word (e.g., *peace*). Elman and McClelland (1988) reasoned that if the basis for the Ganong effect is feedback to the perceptual level, a restored phoneme in that paradigm should have similar *consequences* as an intact phoneme, and in particular, it should drive compensation for coarticulation. They found exactly that result: the boundary of a *tapes-capes* continuum shifted following a segment ambiguous between /s/ and /Σ/ as a function of the lexical bias preceding the ambiguous segment (e.g., *Christma-* or *fooli-*). For the next decade, many regarded this as strong evidence in support of interaction.

However, Pitt and McQueen (1998) explored the hypothesis that the basis for the effect was diphone transitional probabilities (TPs), based on an analysis by Cairns, Shillcock, Chater, and Levy (1995), purportedly showing that Elman and McClelland's lexical contexts were confounded with TP. Under the TP hypothesis, compensation for coarticulation after a segment ambiguous between /s/ and /Σ/ is driven by the higher probability of /s/ after the final vowel of *Christma,* /↔/, than after the final vowel of *fooli*, /I/, and, conversely, the higher probability of /Σ/ after /I/ than after /↔/. Because these transitional probabilities can be viewed as involving sublexical knowledge only, Elman and McClelland's (1986) results would not be proof of lexical influence on sublexical processing. Pitt and McQueen directly tested this hypothesis and found compensation for coarticulation with nonword contexts as a function of TP, but failed to find it in lexical contexts where TP was controlled. For the next several years, this was regarded by many as strong evidence that TP was the basis for "lexically" mediated compensation for coarticulation.

Samuel and Pitt (2003) provided a thorough empirical and acoustic analysis of the paradigm. They reported new studies in which they found lexically mediated compensation for coarticulation with several contexts with opposite lexical and diphone TP biases. They also provided plausible perceptual explanations for the minority of cases where lexically mediated compensation for coarticulation has not been found (e.g., Pitt & McQueen, 1998; and some contexts tested by Samuel and Pitt themselves). Magnuson, McMurray, Tanenhaus, and Aslin (2003a) reported converging evidence as well as a new corpus analysis of transitional probabilities in American English that revealed that not all of Elman and McClelland's lexical contexts were confounded with diphone TP. They also used corpus analyses to show that no particular *n*-phone TP could predict observed lexical effects. Instead, the appropriate TP context seems to be an *n*-phone of dynamic length, where *n* resolves to word length, and thus the knowledge driving mediated compensation for coarticulation seems to be lexical

Further evidence for feedback comes from *selective adaptation to restored phonemes*. Samuel (1997, 2001a, b) has shown that "restored" phonemes (phonemes replaced with noise, but which subjects report hearing in a manner consistent with lexical or larger contexts) can drive the selective adaptation found with fully articulated phonemes. If a segment at one end of a categorical perception continuum is repeated many times, the boundary shifts toward that stimulus, such that a smaller step toward the opposite end of the continuum leads to a change in perception. Restored phonemes have similar (though weaker) effects, suggesting the locus is prelexical.

Norris et al. (2003) added a new wrinkle to the debate. Based on evidence for short-term changes in phonemic categories based on implicit perceptual learning, they acknowledged the need for feedback, but argued that it need not occur on-line. Instead, they make a distinction between on-line feedback (as in interactive models) and *feedback for learning*, although without specifying how feedback for learning is triggered or timed; if it is not to happen during processing, the learning signal must be stored until some opportune "down-time" during which the learning signal may be transmitted. The idea is that since (according to their arguments) on-line feedback can serve no useful role, and since a principled division can be made between on-line and "for learning" feedback in computational models, the most parsimonious account remains an autonomous model with feedback for learning. Norris et al. acknowledge the possibility that feedback might be implemented in such a way that it simultaneously provides on-line and for-learning feedback (see (Mirman), McClelland, & Holt, in press, for just such an implementation, which incorporates Hebbian learning into TRACE), but again, that such an architecture is not necessary; on this view, on-line feedback might exist, but only because it either allows a convenient medium for or is an epiphenomenon of feedback-for-learning.

AQ5

One might argue that in light of the added complexity of post-perceptual decision units in Merge (cf. Samuel, 2001a), the need for feedback to account for perceptual learning, and the ability of a single feedback system to incorporate on-line feedback (accounting for lexical effects on phonemes) and feedback for learning, interaction provides the more parsimonious account. However, given the alternative explanations for the empirical record provided by Norris et al. (2000, 2003), along with their evolving theoretical perspective, there remains room for reasonable disagreement on this debate. Stronger theoretical and empirical cases are required to settle it.

3. SELECTION: HOW IS ACTIVATION REGULATED AND RECOGNITION ACHIEVED?

Once the activation set is specified, a mechanism is needed to evaluate the items in the set and eventually *select* an item for lexical access (and a comprehensive theory must also specify under what conditions selection will fail to occur, e.g., in the case of a nonword input). All current theories assume that a form of competition is required for selection. As a spoken word is heard, multiple lexical items are considered as a function of their phonological similarity to the input and of their frequency of occurrence, or prior probability, and activated lexical items compete for selection. The two key factors we will discuss here are the role of frequency and a sampling of the competition mechanisms proposed under different theories. We will also include a discussion on the issue related to recognizing words in utterances (i.e., the word segmentation issue), as it requires competition among incompatible hypotheses (those that claim the same portion of the input).

3.1. Frequency

It has long been established that words that occur frequently in the language (as reflected by counts of large text corpora) are recognized faster, and more accurately under noisy conditions, than words that occur rarely (e.g., Howes & Solomon, 1951; Savin, 1963). This frequency effect can be couched in Bayesian terms as the impact on perceptual decisions of the prior probability of encountering a given word. The influence of frequency has been instantiated in various ways within theories and models of spoken-word recognition. In search models (e.g., the AUTONOMOUS SEARCH model (Forster, 1989)), word forms are mentally organized into bins, arranged by frequency of occurrence within each bin, with the result that *initial contact* with the lexicon is ordered by frequency. The recognition of a spoken word is viewed as a self-terminating search. The search terminates sooner for high-frequency words, for which a match between the input and a word form can be established early in the search, than for low-frequency words. In localist activation models, which characterize the dominant view in the field, word forms are conceived as independent processing units that accumulate activation proportionally to their match with the incoming signal. In such models, word frequency can directly influence the activation of word units by modulating the units' threshold for response (e.g., the LOGOGEN model (Morton, 1969)), the units' resting (i.e., default) activation (e.g., the COHORT model (Marslen-Wilson, 1987)), the strength of connections between sublexical and lexical representations (MacKay, 1982, 1987), or can act as a post-activation, decision bias, thus acting on *selection* (as in the NAM (Luce, 1986; Luce & Pisoni, 1998; Luce et al., 1990)).

In an attempt to contrast the *initial contact* and *selection* instantiations of frequency, some researchers hypothesized that frequency operating as a decision bias should be observed late, with respect to the onset of spoken input (e.g., Connine, Titone, & Wang, 1993; Goldinger, Luce, & Pisoni, 1989). Although such delay was reported in some studies (Connine et al., 1993), Dahan, Magnuson, and Tanenhaus (2001b) showed that frequency effects could be observed in the earliest moments of lexical processing. They monitored participants' eye movements as they followed spoken instructions to interact with items in a visual display. When fixation proportions over time to low-frequency targets, low-frequency cohorts, high-frequency cohorts and unrelated distractors were compared, Dahan et al. found frequency effects in the earliest signal-driven changes in fixation proportions (within about 200 ms of word onset) – although the magnitude of frequency effects grew as more of a word was heard. Dahan et al. added three frequency mechanisms to TRACE to compare predictions of different proposals for how frequency might be instantiated. Resting level and post-activation bias mechanisms yielded virtually identical predictions (when the post-activation bias was applied continuously, though for it to have a "late" locus it would have to be applied suddenly after a certain amount of bottom-up evidence accrued). A bottom-up connection strength instantiation (in which connections between phonemes and high-frequency words were stronger than those between phonemes and low-frequency words) provided the best fit to the data. This account predicts a continuous effect of frequency, but a gradual one, since the frequency effect depends on the strength of the bottom-up input. The bottom-up connection strength account would also be consistent with learning models in which connection strengths are tuned to prior probabilities through experience.

3.2. Competition

There is now considerable evidence that the recognition of a spoken word is affected by the set of lexical alternatives that are partially compatible with the input. A word that is phonetically similar to few and/or rare other words is recognized more easily than a word similar to many and/or frequent other words, above and beyond effects of the frequency of the word itself (Luce, 1986; Luce & Pisoni, 1998). This indicates that the recognition process does not solely depend on the degree to which the spoken input matches the representation of a given word, but also on the degree to which the input matches the representations of alternative words. All current theories of spoken word recognition acknowledge the need for competition, but differ in the mechanisms they assume accomplishes it. The primary mechanisms are *decision rules* and *direct competition*. We will focus on these, and then turn to a third alternative, *emergent competition*.

Decision rule competition. The original COHORT model (Marslen-Wilson & Welsh, 1978) predicted that the recognition of a spoken word depends on the activation of multiple candidates (the *word-initial cohort*) but only indirectly; the cohort determines the *uniqueness point* of the target word – the point at which the target is the last lexical candidate compatible with the input. The model assumed that the onset of a spoken word activates all word candidates sharing that onset. As more input becomes available, candidates are pruned from the competitor set as soon as they mismatch (e.g., *cat* is removed from *castle*'s cohort when /s/ is heard), until only one candidate remains. Inclusion or exclusion of a candidate from the competitor set was viewed as an all-or-none and frequency-insensitive process. Revisions to the model, prompted by theoretical and empirical arguments (Marslen-Wilson, 1987), changed the mechanism for cohort inclusion and exclusion into a gradient activation process reflecting the degree of evidence for a candidate in the input and its frequency. In this revised model, candidates cannot be described as simply in or out of the cohort. Instead, they are more or less activated, and the criterion for recognition was changed into a decision rule that evaluates a unit's activation level with respect to the activation level of all other units (Marslen-Wilson, 1987, 1993). This, in effect, allows the recognition of a given word to be affected by other candidates' match to the input, but without direct competition between units; any lexical item's activation reflects its goodness of fit to the input. Competition only exists at the level of the decision rule.

A similar mechanism was proposed earlier as part of the NAM developed by Luce and colleagues (Luce, 1986; Luce et al., 1990; Luce & Pisoni, 1998). The model states that the probability of recognizing a given word can be approximated by the ratio of the target word's log frequency to the summed log frequencies of all items in its neighborhood, including the target word; in other words, ease of recognition is predicted to be proportional to the amount of frequency the target contributes to the total frequency of its neighborhood. Computed over large sets of words, this probability rule was shown to account for more unique variance in tasks like lexical decision or naming (about 15%) than any other factor (the next best was target frequency alone, which only accounted for 5%). The NEIGHBORHOOD model stands out among current theories in that it is a formal mathematical model of activation and competition, but not a processing model. It also stands out for its power and simplicity. The frequency-weighted probability rule compactly embodies general principles shared by current theories, as well as the specifics of the neighborhood conception of competitors, and generates precise, testable predictions. Nonetheless, as noted above, the NAM fails to incorporate the dynamics of a spoken word's competition environment.

Direct competition. Connectionist models like TRACE (McClelland & Elman, 1986), SHORTLIST (Norris, 1994), and more recently PARSYN (Luce et al., 2000) assume competition among lexical units via lateral inhibition. Units within the lexical layer (and the phoneme layer, in the case of TRACE and PARSYN) send each other inhibition as a function of their respective activation, which depends on their similarity to the input. For example, upon hearing the input / κατ/ (*cot*), the units *cat* and *cap* would also both be activated; *cat* is more similar to the input than *cap*, and so would be activated more strongly, and send more inhibition to *cap* than viceversa (assuming equal word frequency). The end result is that a lexical item with an activation advantage will eventually suppress its competitors. The recurrent loops created by lateral inhibition in these sorts of models give them temporal dynamics, which allow fine-grained predictions of the activations of targets and competitors over time.

Distinguishing between an implementation of lexical competition in terms of decision rule or lateral inhibition has proven difficult, as they make very similar predictions (Marslen-Wilson, Moss, & van Halen, 1996; see also Bard, 1990). Similar debates are taking place among models of perceptual choice (Usher & McClelland, 2001). Decision-rule competition is arguably a simpler computational mechanism than lateral inhibition. In the decision-rule implementation, the temporal dynamics of candidates' activation can only reflect changes in the evidence supporting each candidate, as the spoken input unfolds over time. By contrast, competition via lateral inhibition predicts temporal dynamics that reflect both the impact of evidence from the input and recurrent loops on candidates' activation. Distinguishing between these two implementations is thus likely to require consideration of lexical activation over time.

Emergent competition. Gaskell and Marslen-Wilson (1997, 1999, 2002) have proposed a distributed architecture, where words are represented by overlapping, distributed patterns of node activation. One portion of these nodes stands for phonological features, while another stands for semantic features. A given word is represented as a pattern of activation among phonological and semantic feature nodes, thus capturing the form and the meaning of that word. When the initial portion of a word is presented to the model, patterns learned by the network that are consistent with the input are simultaneously activated. However, because there is only one substrate for activation–the same set of distributed nodes–the outcome is an activation pattern that *blends* the consistent patterns. Thus, competition takes the form of interference between the patterns associated with candidates consistent with partial input. The activation pattern resulting from processing partial input may be more or less coherent depending on the nature of the information that the nodes encode (phonological vs. semantic) and the number of compatible hypotheses simultaneously considered.

We refer to this as *emergent competition* because the competition dynamics arise from a complex combination of interacting causes. These include intricate patterns of excitatory and inhibitory weights that emerge as a function of the corpus on which a recurrent network is trained, the attractors that form for phonological, semantic, and possibly combinations of inputs and outputs.

The model's distributed architecture makes an intriguing prediction. Although the model assumes the simultaneous activation of all the word candidates that match the input, it also predicts that the resulting pattern of activation does not represent the form or the meaning of any of these candidates individually. Rather, because this activation pattern is a blend, their common features (most often, their shared sounds) are faithfully represented, whereas their divergent features (such as their semantic features, as words that are phonologically similar are not typically semantically related) have been blended; reconstructing the divergent features of word candidates would depend, among other things, on the number of word candidates involved.

Gaskell and Marslen-Wilson (2002) reported data supporting the model's prediction. In particular, they showed that the presentation of a spoken prime that is compatible with several possible candidates (e.g., /κΘπτI/, compatible with a number of candidates, including *captain* and *captive*) does not boost participants' speed at making a lexical decision on a word semantically related to one of the candidates (e.g., commander), suggesting that the semantic representations of the activated phonological forms were blended and not sufficiently distinctive to allow detectable priming. By contrast, the presentation of a spoken prime that is compatible with only one possible candidate (e.g., /γμ↔/, only AQ6
compatible with *garment* [British English pronunciation]) did facilitate processing of a word semantically related to this candidate (e.g., attire). This result can be accounted for by the distributed architecture assumed by Gaskell and Marslen-Wilson's model because the pattern of activation in the semantic feature nodes becomes less coherent as more candidates are considered and more heterogeneous patterns (associated with form-overlapping candidates with unrelated meanings) participate in the blend.

Models with localist representations could also account for this result. We are unaware of any current, implemented model that could do so without modification, but the general principles of, e.g., interactive activation are consistent with the result. An explanation parallel to that of Gaskell's and Marslen-Wilson's is that the larger the phonological competitor set is, the weaker the activation that each of their semantic representation receives. The phonological competitors initially receive equivalent support from the phonological input (*mutatis mutandis* for differences in frequency, etc.). As long as no phonological representations are strongly favored by the bottom-up input, however, their corresponding semantic representations receive too little activation to be detected via priming. An analogous mechanism exists in ARTWORD (Grossberg & Myers, 2000), where a perceptual resonance (assumed to lead to conscious perception) is established only once the level of activation of one candidate (or "chunk") has sufficiently overcome that of its competitors.

3.3. Word Segmentation in Continuous Speech: Competition Across Word Boundaries

A spoken utterance cannot easily be segmented into the words that compose it because boundaries between words are not reliably marked in the acoustic signal, and have often been blurred through phonological phenomena such as coarticulation and resyllabification. This is not to say that word boundaries are never acoustically marked. For instance, silent pauses between phrases mark the boundaries of the words that appear at the edges of these phrases. In fact, an extensive literature has demonstrated that listeners make use of word-boundary cues when present (phonotactic cues: McQueen, 1998; prosodic cues: Salverda et al., 2003; phonetic cues: Quené, 1992, 1993; Gow & Gordon, 1995). What this literature has shown is that word-boundary cues are used as a source of evidence supporting word candidates that are consistent with the hypothesized word boundary, and not used prelexically, to chunk the signal into words *before* initiating contact with the lexicon, as had been previously proposed (e.g., Cutler, 1990).

Because word boundary cues are probabilistic at best, and because words tend to share many of their components with other words, multiple words are consistent with virtually any portion of an utterance. For example, McQueen et al. (1995) established that 84% of English polysyllabic words contain at least one shorter embedded word (e.g., *ham* in *hamster*, or *bone* in *trombone*). This lexical ambiguity sometimes applies across word boundaries, as in *ship inquiry*, where (in British English) *shipping* matches *ship* and the initial portion of *inquiry*. Thus, competition among word candidates that start at different points in time is required. As mentioned earlier, TRACE models inter-word competition by assuming that all word units that overlap in time, i.e., competing for the same portion of the input, inhibit one another. Because a unit representing the same word is replicated many times over time/space, a given word unit can become activated as soon as the input provides some evidence supporting it, regardless of where in time the information appears. For instance, after the sequence /ΣIπIN/ (the initial portion of the phrase *ship inquiry*), *inquiry* can start receiving activation from the input, and eventually be recognized, even though *shipping* is already strongly activated. Note that some words can compete even when they do not share any segments. In the example above, the candidate *shipment* competes with *inquiry* because both are competing for the same portion of the input. Thus, TRACE solves the problem of segmenting words out of a continuous spoken input by using the same mechanism it uses to segment a coarticulated signal into a sequence of phonemic units.

Alternatives to TRACE's solution to word segmentation and recognition have been proposed. Norris (1994) criticized the multiple replications of the lexical network in TRACE. He developed SHORTLIST, a model in which a limited set of candidates that are most activated by (i.e., consistent with) the input is compiled. The model consists of two components. A lexical search network, implemented as a simple dictionary lookup, provides a list of the best matches to the input at each phoneme position. The second component is a competition network including as many as the top 30 candidates aligned with each input position (SHORTLIST is often described as allowing a maximum of 30 words to enter the competition network, but this is inaccurate; D. Norris, personal communication). Items selected for each shortlist compete with one another proportionally to the number of sounds they share in an interactive activation network. Items in different shortlists also compete *if* they overlap. For example, given the input *ship inquiry*, *ship* and *shipping* will enter the shortlist aligned with the first phoneme. *Inquiry* will eventually dominate the shortlist aligned with the fourth phoneme, i.e., after *ship*, and will inhibit *shipping*, because the two overlap in input positions 4 and 5, but it will not inhibit *ship*, since it does not overlap with *ship*. Thus, *ship* and *inquiry* create pressure for a parse into nonoverlapping words, and eventually inhibit *shipping* sufficiently to allow *ship* to be recognized. The selection–competition cycle repeats itself as input is presented to the model. At each time step, a new lexical search is done for *every position encountered so far*. The composition of the shortlist changes dynamically as spoken input becomes available, with some candidates dropping and being replaced by new candidates, depending on bottom-up match/mismatch scores from the lexical search network and inhibition within the competition network.

Despite the important computational economy offered by establishing the competitor set in a dynamical fashion, compared to a hard-wired manner as in TRACE, SHORTLIST also has several limitations. First, the lexical search mechanism is called recursively–a new search is done at each position as each new phoneme is heard. If the lexical search were implemented as a recurrent network, this would require one copy of the lexical network for each phoneme position, and so the model would require the same number of nodes as TRACE, plus those used in the shortlists (but would use many fewer *connections*). Second, the biological plausibility of the dynamic programming required by SHORTLIST must be addressed (cf. Protopappas, 1999). Finally, it has yet to be shown that SHORTLIST can account for the broad range of data TRACE can.

ARTWORD (Grossberg & Myers, 2000) is a model specifically designed to account for the dynamics of inter-word competition and how later-arriving information can modulate the perception of earlier occurring speech. In this model, the spoken input activates sensory features. Activation of these features is transformed into a sequence of items in working memory. The sequential order of these items is encoded by a gradient of activity within the representation (with the most active item representations corresponding to the most recent event). The activity pattern in working memory in turn activates "list chunks" that match the active items *and* their order. List chunks consist of unitized linguistic units (e.g., phonemes, syllables, words). Activated chunks compete with one another, proportionally to their level of activation and to the number of items they compete for. Once an activated list chunk reaches an activation threshold, it sends back activation to the consistent items in working memory, and inhibition to inconsistent items.

The excitatory loop between list chunks and items in working memory corresponds to a process known as *resonance*. In Grossberg and Myers's (2000) own words, "when listeners perceive fluent speech, a wave of resonant activity plays across the working memory, binding the phonemic items into larger language units and raising them into the listener's conscious perception" (p. 738). Thus, in this model, recognizing a spoken word can be described as having associated a given linguistic interpretation to a portion of speech represented in working memory, where time is encoded.

The dynamics of the resonance wave is the major factor that determines how continuous speech is perceived as a succession of segmented and unitized word units. First, the model includes two reset mechanisms that can terminate one resonance to allow for the next one to be initiated (see Grossberg, Boardman, & Cohen, 1997, for more details). Thus, the perception of a multi-word utterance can be described as a sequence of resonance waves. Second, because of competition among activated chunks, ARTWORD accounts for recognition despite the activation of multiple candidates at various points in the signal. Third, the model allows for later-arriving information to modify the resonance wave by *resonant transfer*: The resonance associated with a short word (e.g., *ham*) can be transferred to a longer one (e.g., *hamster*) as the second syllable of the word *hamster* is processed. Finally and critically, ARTWORD can account for the impact of some word-boundary cues (such as segmental lengthening, e.g., Salverda et al., 2003) without invoking additional mechanisms. Indeed, a resonance transfer can only occur within a very limited, speech-rate-dependent time window. Thus, if the first sounds of the second syllable of *hamster* are delayed (because of lengthening of the last sounds of *ham*, a silent pause, or lengthening of the sound following *ham*), the resonance established between the word chunk *ham* and items in working memory may have been reset, and the items' activation fallen to low activation levels. No resonance transfer is then possible, and listeners will perceive the word *ham* followed by another word starting with the sounds /st/. This is consistent with Salverda et al.'s results, showing that long /ham/ syllables tend to be interpreted as monosyllabic words.

4. INTEGRATION: WHEN AND HOW IS CONTEXT INTEGRATED?

Words occur embedded in a larger context, most often in a sentence. There exists a tight interdependency between a given word and its sentential context. A word contributes to the meaning of the sentence, but the contribution of a word to the meaning of the sentence also rests on the sentence itself.

Most of the empirical work examining the interaction between a word and its sentential context has focused on the possible constraint that the context may impose on the set of word candidates compatible with the spoken input. Initial studies suggested a late impact of context. For example, Tanenhaus, Seidenberg, and Leiman (1979; see also Swinney, 1979) presented listeners with auditory sentences that were biased toward one sense of a homophone (e.g., *she held the rose* vs. *they all rose*), and then used visual lexical decision to probe semantic activation. They found statistically equivalent priming for associates of both senses (e.g., *flower* and *stand*) immediately after homophone offset, but only found reliable priming for the context-appropriate sense 250 ms later. This was interpreted as evidence for context-free initial lexical activation, quickly followed by an integration stage where word interpretations incompatible with the context are rejected. Similar conclusions were reached by Zwitserlood (1989), who reported evidence for the early activation of the meaning of all words compatible with the initial sounds of a spoken word, regardless of the context.

However, Shillcock and Bard (1993) tested the hypothesis that the Tanenhaus et al. contexts contained very weak biases (other form classes besides nouns or verbs could have been heard at the homophone position, and the contexts at best biased listeners toward thousands of nouns vs. thousands of verbs). They used contexts that had been experimentally established as biased towards a single item – the closed class word, *would* (*John said he didn't want to do the job but his brother would, as I later found out*) – or towards a large number of items: (*John said he didn't want to do the job with his brother's wood, as I later found out*). In the closed-class case, they found no evidence of priming of *wood*; its associate, *timber*, was not primed even if they probed prior to the offset of *would*. This suggests that top-down context *can* affect early stages of word recognition, but that top-down information is generally given much less weight than bottom-up, and is proportional to prior probability: the more narrowly constraining the top-down information is, the greater the impact it may have on early moments of processing (see Dahan, Swingley, Tanenhaus, & Magnuson, 2000, who report evidence for the early impact of determiners marked for grammatical gender on the recognition of subsequent spoken nouns in French).

Generally speaking, theories of spoken word recognition have remained agnostic about the integration of sensory information with higher level context. Notable exceptions to this are the three versions of the COHORT model. In the original COHORT model, top-down knowledge (e.g., semantic context) played an active role throughout selection, allowing recognition prior to the uniqueness point for words strongly supported by context. It also had the potential to guide initial contact, by preventing a highly inconsistent item from entering the recognition cohort. In the revised COHORT model, in light of intuitive and empirical evidence that clearly articulated words that have low probability in a particular context are still clearly perceived, context no longer affected initial contact (i.e., could no longer exclude an item from entering the cohort despite strong bottom-up support). Instead, context was viewed as acting on a set of candidates first established on the basis of sensory information only. The model argued in favor of a context-free, initial activation stage. The most recent version of the model, the DISTRIBUTED COHORT model, departs from this stance by assuming no division between initial contact and selection. Semantic features are an integral part of lexical representations, and thus semantic and phonological knowledge are simultaneously activated by bottom-up input. This last instantiation, by renouncing the theoretical processing division between form and meaning, is compatible with findings of a continuous integration of different sources of evidence in order to ultimately derive an interpretation of the spoken input.

5. AVENUES FOR PROGRESS

The three most crucial developments for theories of spoken-word recognition, as argued throughout this chapter, are (1) increasing evidence that the input to spoken word recognition retains much if not all of the surface detail of utterances; (2) evidence that language representations are not static but instead are subject to constant change; and (3) the emergence of theoretical frameworks that deny the existence of distinct stages corresponding to speech perception, spoken word recognition, sentence processing, and beyond – and empirical support for these theories. These developments may herald a radical reconceptualization of spoken word recognition and language processing in general, if not an all-out paradigm shift.

There are two sets of findings that compellingly demonstrate that the input to lexical access is not limited to an abstract phonemic code. The first (reviewed briefly in our introduction) is evidence that fine-grained phonetic detail affects the time course of lexical activation and competition (Andruski et al., 1994; Davis et al., 2002; Salverda et al., 2003). The second (reviewed in Section 2) is evidence that even (putatively) non-linguistic surface detail, such as talker sex or even more fine-grained talker characteristics, is preserved in memory for spoken language (Goldinger, 1996). The fact that such detail not only affects memory but also word recognition motivates exemplar theories like Goldinger's (1998) episodic lexicon theory, in which the basis for lexical (and potentially lower and higher levels of representation) categories are clusters of memory traces of, essentially, raw speech "episodes" that preserve all surface detail. On such a view, each new memory trace has the potential to change the "category" with which it is clustered, making exemplar theories compatible with recent evidence that short-term changes in phonotactic probabilities quickly influence production (Dell, Reed, Adams, & Meyer, 2000) and comprehension (Onishi, Chambers, & Fisher, 2002). These rapid changes in lexical production and processing challenge the frequent, if implicit, assumption that the adult phonological and lexical knowledge is more or less fixed.

These developments pose significant challenges to theories of spoken word recognition and spoken language processing in general. They point to a system in which there may be distinct levels of representation (given the cognitive economies of compositionality and generativity afforded by, e.g., phonemes and words), but also parallel episodic representations that are less abstract, and without discrete stages corresponding to the descriptive levels of speech perception, word recognition, sentence processing, and so on.

As mentioned earlier, Goldinger and Azuma's (2003) appeal to adaptive resonance (e.g., Grossberg, 2003) as a potentially unifying framework capable of incorporating learning, sublexical and lexical effects as well as the principles of episodic lexicon theory, appears to hold substantial promise.

However, integrating this view with the processing of actual speech or a close analog, remains a significant challenge. While the ARTSTREAM model (Grossberg et al., 2004) has demonstrated the potential of the ART framework to process the speech signal itself, it has not yet been extended to contact with phonemic or lexical forms. Plaut and Kello (1999) provided another framework with significant promise, in which close analogs of the speech signal are used, and phonological and semantic representations are treated within perception and production, as well as development.

Integrating (descriptive) levels of speech perception and word recognition upwards also remains as a significant challenge. Theories of sentence processing in the constraint-based framework have long blurred the boundary between lexical access and sentence processing (e.g., MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell & Kim, 1998), assuming that lexical representations include not just phonological and semantic knowledge, but also specify the syntactic relations in which a lexical item can participate. Evidence that lexical access and sentence processing are constrained in an immediate and continuous fashion by nonlinguistic context – such as the actions afforded to the listener by combinations of objects and instruments (Chambers, Magnuson, & Tanenhaus, 2004), or even affordances available to interlocuters (Hanna & Tanenhaus, 2003) – demands that we scale our theories up and integrate them with sentence, discourse, and general cognitive processes.

We began this chapter by contrasting the modal, narrow view of spoken word recognition (as the mapping from phonemes to sound forms that provide access to the lexicon) with a broad view, encompassing the speech signal, the word level, and higher levels of structure and representation. The broad view is supported by growing evidence for continuous effects of subphonemic information at the lexical level and beyond on the one hand, and immediate integration and interaction between descriptively low and high levels of linguistic representation and even non-linguistic affordances of physical objects (Chambers et al., 2004) on the other. Our view is that significant progress in understanding spoken word recognition, and language processing more generally, will require stretching (or possibly abandoning) current theories and models to accommodate the broad view of language processing.

ACKNOWLEDGEMENTS

We thank Dan Swingley and Dan Mirman for comments and discussions that greatly improved this chapter. Preparation of this chapter was supported in part by the National Science Foundation under Grant No. 0433567 to DD, and by National Institute on Deafness and Other Communication Disorders Grant DC-005765 to JSM.

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 & Language*, *38*, 419–439.

Altmann, G. T. M. (1999). Thematic role assignment in context. *Journal of Memory & Language*, *41*, 124–145.

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

Bagley, W. C. (1900). The apperception of the spoken sentence: A study in the psychology of language. *American Journal of Psychology*, *12*, 80–130. AQ7

Bailey, T. M., & Hahn, U. (2005). Phoneme similarity and confusability. *Journal of Memory and Language*, *52*, 339–362.

Bard, E. G. (1990). Competition, lateral inhibition, and frequency: Comments on the papers of Frauenfelder & Peeters, Marslen-Wilson, and others. In: G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 185–210). Cambridge, MA: MIT Press.

Bertelson, P., & de Gelder, B. (1989). Learning about reading from illiterates. In: M. Galaburda, & M. Albert (Eds), *From reading to neurons* (pp. 1–23). Cambridge, MA: MIT Press.

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. K. (2004). Reference resolution in the wild: Circumscription of referential domains by naive participants during an interactive problem solving task. In: J. C. Trueswell & M. K. Tanenhaus (Eds), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*. Cambridge, MA: MIT Press.

Cairns, P., Shillock, R., Chater, N. & Levy, J. (1995) Bottom-up connectionist modelling of speech. In: J. P. Levy, D. Bairaktaris, J. A. Bullinaria, & P. Cairns (Eds). *Connectionist models of memory and language*, University College London Press. AQ8

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

Clarke, C. M., and Garrett, M. F. (2004) Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America*, *116*, 3647–3658.

Cluff, M. S., & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception & Performance*, *16*, 551–563.

Cole, R. A. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception & Psychophysics*, *13*, 153–156.

Cole, R. A., & Rudnicky, A. I. (1983). What's new in speech perception? The research and ideas of William Chandler Bagley, 1874–1946. *Psychological Review*, *90*, 94–101.

Connine, C. M. (2004). It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition. *Psychonomic Bulletin & Review*, *11*, 1084–1089.

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

Cutler, A. (1990). Exploiting prosodic probabilities in speech segmentation. In: G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (pp. 105–121). Cambridge: MIT Press.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language & Speech*, *40*, 141–201.

Cutler, A., & Norris, D. (1979). Monitoring sentence comprehension. In: W. E. Cooper, & E. C. T. Walker (Eds), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett.* Hillsdale: Erlbaum.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001a). 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. (2001b). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*, 507–534.

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in french. *Journal of Memory and Language*, *42*, 465–480.

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, & Cognition*, *30*, 498–513.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*, 222–241.

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 & Performance*, *28*, 218–244.

Deelman, T., & Connine, C. M. (2001). Missing information in spoken word recognition: Nonrelased stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 656–663.

Dell, G.S., Reed, K.D., Adams, D.R. & Meyer, A.S. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *26*, 1355–1367.

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception & Performance*, *23*, 914–927.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*, 224–238.

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, *8*, 301–306.

Elman, J. L., & McClelland, J. L. (1986). Exploiting lawful variability in the speech wave. In: J. S. Perkell, & D. H. Klatt (Eds), *Invariance and variability in speech processes*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Forster, K. I. (1989). Basic issues in lexical processing. In: W. Marslen-Wilson (Ed.), *Lexical representation and process*. Cambridge, MA: MIT Press.

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception and Psychophysics*, *36*, 359–368.

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

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 & 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., Hare, M., & Marslen-Wilson, W. D. (1995). A connectionist model of phonological representation in speech perception. *Cognitive Science*, *19*, 407–439.

Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception & Performance*, *22*, 144–158.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language & Cognitive Processes. Special Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives on the Lexicon*, *12*, 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. (2001). Lexical ambiguity resolution and spoken word recognition: Bridging the gap. *Journal of Memory & Language*, *44*, 325–349.

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

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166–1183.

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

Goldinger, S.D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*, 305–320.

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

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

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

Gow, D. W., Jr. (2003a). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & 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.

Gow, D. W., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception & Performance*, *21*, 344–359.

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *14*, 421–433.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, *31*, 423–445.

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception & Performance*, *23*, 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*, 511–536.

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

Hahn, U., & Bailey, T. M. (2004). What makes words sound similar? *Cognition* (in press). **AQ9**

Hanna, J.E. & Tanenhaus, M.K. (2003). Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science*, *28*, 105–115.

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

Hintzman, D. L., Block, R. A., & Inskeep, N. R. (1972). Memory for mode of input. *Journal of Verbal Learning and Verbal Behavior*, *11*, 741–749.

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

Joos, M. (1948). *Acoustic phonetics*. Baltimore, MD: Linguistic Society of America.

Kjelgaard, M. M., & Speer, S. R. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory & Language*, *40*, 153–194.

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

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

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

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

Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology*, *18*, 201–212.

LoCasto, P., & Connine, C. M. (2002). Rule-governed missing information in spoken word recognition: Schwa vowel deletion. *Perception and Psychophysics*, *64*, 208–219.

Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics*, *39*, 155–158.

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., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition*, *26*, 708–715.

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

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In: G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 122–147). Cambridge, MA: MIT Press.

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

MacKay, D. G. (1982). The problems of flexibility, fluency, and speed-accuracy trade-off in skilled behavior. *Psychological Review*, *89*, 483–506.

MacKay, D. G. (1987). *The organization of perception and action: A theory for language and other cognitive skills*. New York: Springer.

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.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [∑]-[s] distinction. *Perception & Psychophysics*, *28*, 213–228.

Marslen-Wilson, W. (1984). Function and process in spoken word recognition. In: H. Bouma & D. G. Bouwhis (Eds), *Attention and performance x: Control of language processes* (pp. 125–150). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Marslen-Wilson, W. (1989). Access and integration: Projecting sound onto meaning. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 3–24). Cambridge, MA: MIT Press.

Marslen-Wilson, W.D. (1990). Activation, competition, and frequency in lexical access. In: G.T.M. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (pp.148–172). Cambridge, MA: MIT Press.

Marslen-Wilson, W. (1993). Issues of process and representation in lexical access. In: G. T. M. Altmann, & R. Shillcock (Eds), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting*, pp. 187–210. Mahwah: Erlbaum.

Marslen-Wilson, W., Moss, H. E., & van Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception & Performance*, *22*, 1376–1392.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representations and process in lexical access: Words, phonemes, 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., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception & Performance*, *15*, 576–585.

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

McClelland, J. L., St John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language & Cognitive Processes. Special Parsing and Interpretation*, *4*, SI287–SI335.

McLennan, C. T., Luce, P. A., & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*, 539–553.

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

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory & Language*, *39*, 21–46.

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., Norris, D., & Cutler, A. (1999). Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception & Performance*, *25*, 1363–1389.

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

Mirman, D., McClelland, J. L., & Holt, L. L. (2005). Computational and behavioral investigations of lexically induced delays in phoneme recognition. *Journal of Memory & Language*, *52*, 416–435.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*, 165–178.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, *85*, 365–378.

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

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

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

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. In: K. Johnson, & J. W. Mullennix (Eds), *Talker variability in speech processing* (pp. 109–132). San Diego, CA: Academic Press.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In: Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds), *Speech perception, speech production, and linguistic structure* (pp. 113–134). Tokyo: OHM.

Onishi, K. H., Chambers, K. E., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition*, *83*, B13–B23.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *19*, 309–328.

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

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, *25*, 21–52.

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

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. 381–416). Mahwah, NJ: Lawrence Erlbaum Associates.

Protopapas, A. (1999). Connectionist modeling of speech perception. *Psychological Bulletin*, *125*, 410–436.

Quené, H. (1992). Durational cues for word segmentation in dutch. *Journal of Phonetics*, *20*, 331–350.

Quené H. (1993). Segment durations and accent as cues to word segmentation in dutch. *Journal of the Acoustical Society of America*, *94*, 2027–2035.

Remez, R. E. (2005). The perceptual organization of speech. In: D. B. Pisoni, R. E. Remez (Eds), *The Handbook of Speech Perception*, (pp. 28–50). Oxford: Blackwell.

Rubin, P., Turvey, M. T., & Van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception & Psychophysics*, *19*, 394–398.

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

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

Samuel, A. G. (2001b). Some empirical tests of merge's architecture. *Language & Cognitive Processes*, *16*, 709–714.

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

Savin, H.B. (1963). Word frequency effects and errors in the perception of speech. *Journal of the Acoustical Society of America*, *35*, 200–206.

Scharenborg, O., ten Bosch, L., Boves, L., & Norris, D. (2003). Bridging automatic speech recognition and psycholinguistics: Extending shortlist to an end-to-end model of human speech recognition (l). *Journal of the Acoustical Society of America*, *114*, 3032–3035.

Shankweiler, D., & Fowler, A. E. (2004). Questions people ask about the role of phonological processes in learning to read. *Reading and Writing: An Interdisciplinary Journal*, *17*, 483–515.

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. 163–183). Mahwah, NJ: Lawrence Erlbaum Associates.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition: I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, *96*, 1314–1324.

Speer, S., & Blodgett, A. (2006). Prosody. In: M. A. Gernsbacher, & M. Traxler (Eds), *Handbook of Psycholinguistics*. Amsterdam: Elsevier.

Sumner, M., & Samuel, A. G. (2005). Perception and representation of regular variation: The case of final /t/. *Journal of Memory and Language*, *52*, 322–338.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning & 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 & 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*, 1632–1634.

Trueswell, J.C., & Kim, A.E. (1998). How to prune a garden-path by nipping it in the bud: Fast-priming of verb argument structures. *Journal of Memory and Language*, *39*, 102–123.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*, 550–592.

Utman, J. A., Blumstein, S. E., & Burton, M. W. (2000). Effects of subphonetic and syllable structure variation on word recognition. *Perception and Psychophysics*, *62*, 1297–1311.

van Ooijen, B. (1996). Vowel mutability and lexical selection in english: Evidence from a word reconstruction task. *Memory and Cognition*, *24*(5), 573–583.

Warren, R. M. (1970). Restoration of missing speech sounds. *Science*, *167*, 392–393.

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics*, *35*, 49–64.

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

AUTHOR QUERY FORM

ELSEVIER

Handbook of Psycholinguistics

JOURNAL TITLE:	**HP-TRAXLER**
ARTICLE NO:	**Ch008**

Queries and / or remarks

Query No	Details required	Author's response
AQ1	In Hintzman et al. (1972), Hintzman is spelt as Hintzman and Hinztman. Please confirm the correct spelling.	
AQ2	In Hintzman (1986), Hintzman is spelt as Hintzman and Hinztman. Please confirm the correct spelling.	
AQ3	Please confirm whether Dahan et al. (2001) corresponds to 2001a or 2001b.	
AQ4	Nygaard et al. (1994), Mann & Repp (1981), Mirman (in press), Tanenhaus, Seidenberg, and Leiman (1979) are not listed in Reference list.	
AQ5	Please provide the year of publication in Mirman et al. (in press). Please check whether Mirman et al. (in press) refers to Mirman et al. (2005).	
AQ6	Please check whether any symbol is missing in between 'γ' and 'μ' in the sentence "By contrast, the presentation ... to this candidate (e.g., attire)."	
AQ7	Bagley (1900), Cole & Rudnicky (1983), Kuhn (1962), Mann & Repp (1980), Mirman et al. (2005), Tanenhaus, Leiman, and Seidenberg (1979) are not cited in the text.	
AQ8	Please provide the place of publication in Cairns et al. (1995).	
AQ9	Please provide the vol. no. and page range in Hahn & Bailey (2004).	

AUTHOR QUERY FORM

ELSEVIER

Handbook of Psycholinguistics

JOURNAL TITLE:	**HP-TRAXLER**
ARTICLE NO:	**Ch008**

Queries and / or remarks

Query No	Details required	Author's response