MAGNUSON, J. S., MIRMAN, D., & HARRIS, H. D. (2012). COMPUTATIONAL MODELS OF SPOKEN WORD RECOGNITION. IN M. SPIVEY, K. MCRAE, & M. JOANISSE (EDS.), *THE CAMBRIDGE HANDBOOK OF PSYCHOLINGUISTICS*. (pp. 76-103). CAMBRIDGE UNIVERSITY PRESS.

Computational models of spoken word recognition

James S. Magnuson
Daniel Mirman
University of Connecticut and Haskins Laboratories

Harlan D. Harris
New York University

1. Preliminaries

A broad distinction can be drawn in psycholinguistics between research focused on how input signals activate representations of linguistic forms, and how linguistic forms are used to access or construct conceptual representations. Words lie at the junction, but do more than simply provide an interface between signals and higher-level structures. Theories in psycholinguistics (e.g., MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell & Tanenhaus, 1994) and linguistics (e.g., Pustejovsky, 1995) have ascribed increasing syntactic and semantic knowledge and function to the lexical level. This makes theories of spoken word recognition (SWR) key in explaining not just how word forms are recognized, but also in understanding levels upstream (sublexical) and downstream (conceptual, sentential, etc.). While theories of SWR typically take the narrow focus of mapping from phonemes to sound patterns of words, a growing body of empirical results (consistent with the increasing role of the lexicon in linguistic and psycholinguistic theory) suggests that SWR is not so neatly compartmentalized. For example, subphonemic details in the speech signal affect lexical activation (Andruski, Blumstein, & Burton, 1994; Davis, Marslen-Wilson, & Gaskell, 2002; Salverda, Dahan, & McQueen, 2003), revealing that sublexical details are preserved at least to the level of lexical access. Lexical context appears to influence sublexical perception directly (Elman & McClelland, 1988; Samuel, 1981; but see discussion of controversies on this point below), and syntactic context similarly influences lexical activation (Shillcock & Bard, 1993; van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005).

Determining what representations are active during any cognitive process is difficult, since many of those representations may no longer be active by the end of the process. The problem is compounded by the nature of the speech signal. The transient acoustic events that make up spoken words must be mapped rapidly onto words in memory, within the limits of echoic and working memory. SWR is further complicated by the many-to-many mapping between acoustics and linguistic categories (Fowler & Magnuson, this volume) and the absence of invariant cues to word boundaries (Samuel & Sumner, this volume), placing speech perception and SWR among the most challenging problems in cognitive science.

In tackling these problems, theories of SWR generally agree on three principles (Dahan & Magnuson, 2006). First, as a word is heard, multiple lexical representations are activated. Second, activation depends on degree of fit between a lexical item and the incoming speech, and prior probability (frequency of occurrence). Third, recognition is guided by competition among activated representations. Each principle is quite general, and allows for considerable variation in

specifics. Theories differ particularly in their similarity metrics and/or bottom-up activation mechanisms (which determine degree of fit), information flow (e.g., only bottom-up or top-down as well), and the nature of the competition mechanisms they assume.

Different assumptions about these principles lead to different predictions about word recognition. Current theories are generally guided by *computational models*, which minimally include *mathematical*, *verbal-algorithmic,* and *simulation* models.[i] In the next section, we will give one example of each of the first two types, and then review several simulation models, introducing additional distinctions among model types as needed. Our review necessarily will be brief and selective, with models chosen to illustrate approaches and principles. For more comprehensive reviews, see Protopapas (1999) and Ellis and Humphreys (1999). We will then review a recent debate in SWR that hinges on subtle predictions that follow from computational models but have proved elusive in empirical tests. The debate provides useful illustrations of principles of model testing and comparison. We will close the chapter with a discussion of what we see as the most pressing issues for making progress in theories of SWR, and the most promising current modeling approaches.

2. A selective review of SWR models

2.1. Mathematical models

The most influential *mathematical* model of SWR is the *Neighborhood Activation Model* (NAM; Luce, 1986; Luce & Pisoni, 1998), which crystallizes the three key SWR principles reviewed above into a simple, but powerful, mathematical form. NAM is also the only SWR model able to generate item-specific and pair-wise competition predictions for thousands of words *easily*. Luce and Pisoni discuss potential connections with simulation models like TRACE (see Section 2.3) and have proposed PARSYN as a simulating instantiation of NAM (Luce, Goldinger, Auer, & Vitevitch, 2000), but NAM itself does not specify any algorithms or mechanisms. Rather, it combines general principles and constraints on SWR into a mathematical form that predicts relative ease of lexical access.[ii]

This simplicity also places NAM at the fundamentalist end of a *fundamentalist-realist continuum* of models (Kello & Plaut, 2003). Fundamentalist models isolate key theoretical assumptions and implement them with as little baggage as possible, with the goal of making transparent tests of the assumptions. Realist models build in as much detail as possible, with the goal of accounting for a broad and deep range of phenomena, often with the goal of seeing whether the complexity of the model engenders emergence of unexpected (positive or negative) behavior.

How does NAM formalize the three core principles of SWR? First, it addresses multiple activation and similarity with a *global similarity metric* that specifies which words will be activated as a word is heard, and how strongly they will be activated. The most familiar NAM metric uses a *one-phoneme "DAS" (deletion, addition, or substitution) threshold*: words are neighbors if they differ by no more than one phoneme, whether by deletion (*cat: at*), addition (*cat: scat, cast, cattle*), or substitution (*cat: bat, cot, cab*). More subtle metrics based on empirical measures of sublexical similarity (e.g., perceptual confusion data) can also be used to compute pair-wise positional similarity over all words in the lexicon (where overall similarity of two words is the product of phoneme-by-phoneme similarities). While the more complex metrics do make distinct predictions, such as the priming of *veer* by *bull* (given high similarity at each phoneme; Luce et al., 2000), the two metrics make sufficiently similar predictions that the one-phoneme metric is most frequently used.

Once the neighborhood of a word is defined (or computed, in the case of graded similarity metrics), a word's frequency-weighted neighborhood probability can be computed. We present a slightly modified version of the Luce and Pisoni (1998) form in Equation 1, where $FWNP_t$ is the frequency-weighted neighborhood probability of target word t, f_t is the prior probability (typically, the log frequency of occurrence per million words in a corpus) of a target word t, and s_t is the similarity of the target to itself (which may approach but not equal 1.0 in some versions of the metric -- e.g., when similarity is based on phonemic confusion probabilities). In the denominator, for every word, w, in the lexicon (including the target), f_w is the frequency of word w, and s_{wt} is the similarity of word w with target t. Note that if a threshold rule (like the DAS rule) is not used to define neighbors, the set of potential neighbors includes every word in the lexicon, though many words will have similarities to t near 0. Note also that denominator includes the target, t; even when a threshold is used, t will be a neighbor of itself.

$$FWNP_t = \frac{f_t s_t}{\sum f_w s_{wt}} \tag{1}$$

This is the most general form of the rule. When the DAS definition of neighbor is used, we can simplify further by dropping the s terms, as items either have similarity of 1.0 (meets DAS definition of neighbor) or 0.0 (not a DAS neighbor).

NAM addresses *prior probability* by weighting each neighbor in the metric by its log frequency. NAM addresses *competition* indirectly, with a choice rule that approximates lexical competition. Ease of recognition of a target word is predicted by the ratio of its log frequency to the sum of all other words' similarities to the target (0 or 1 for the DAS rule) weighted by each item's log frequency. Since neighborhood density (summed frequency-weighted neighbor similarities) includes the target (with self-similarity of 1), frequency-weighted neighborhood probability can be stated more simply as *the proportion of the neighborhood frequency contributed by the target word*. NAM predicts that if two words are matched on neighborhood, the one with higher frequency will be recognized more quickly, because it contributes a larger portion of its neighborhood density. If two words are matched on frequency, the one with lower neighborhood density will be recognized more quickly, again because that word's frequency represents a greater proportion of its neighborhood density. Note that the temporal grain size of the model is lexical – it simply predicts the recognition facility of entire words, and does not predict sublexical processing details.

This simple model is surprisingly powerful. NAM accounts for about 15% of the variance (beyond word frequency alone) in tasks like lexical decision and naming (Luce, 1986; Luce & Pisoni, 1998). The next best predictor is frequency alone – which only accounts for about 5% of the variance. Significant effects are commonly found in factorial manipulations of neighborhood density, and again, the complex similarity metric makes surprising pair-wise priming predictions that have been borne out empirically (Luce et al., 2000). NAM has had a large impact on theories and the practice of SWR research (studies of SWR now commonly control neighborhood density).

NAM can be considered a general framework for choice models of SWR, or as a specific, testable model when paired with a particular metric. While the model is strongly associated with the metrics used by Luce and colleagues and the competitor set predictions that follow, using other similarity metrics in the NAM framework would be an excellent strategy for making further progress on identifying general constraints on SWR.

2.2. Verbal-algorithmic models

In *verbal-algorithmic* models, predictions that follow from theoretical assumptions are *described* as an ordered series of processes or computations. The preeminent example in SWR is the *Cohort Model* developed by Marslen-Wilson and colleagues (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978). The Cohort Model illustrates the power of a well-specified verbal-algorithmic model, as it makes many testable predictions and paved the way for the simulation models we describe next. Cohort differs from NAM in three key ways. First, of course, it is a verbal-algorithmic formulation of processing mechanisms that could support SWR rather than a mathematical formulation of general principles. Second, algorithmic choices lead to a similarity metric that differs considerably from NAM's. Third, it grapples explicitly with challenges of processing the speech signal over time, which allows it to generate qualitative time-course predictions and address segmentation of fluent speech.

The original Cohort model was formulated to account for constraints that emerged primarily from experiments that revealed that SWR can occur remarkably early, prior to word offset, depending on possible competitors in the lexicon and higher-level context (Marslen-Wilson & Welsh, 1978). Cohort built on the activation metaphors introduced in Morton's (1969) Logogen theory and broke SWR into three stages: *access* (initial contact of bottom-up perceptual input with lexical representations), *selection* (winnowing the activation *cohort*), and *integration* (retrieving syntactic and semantic properties of a selected word and checking compatibility with higher levels of processing). The key theoretical constraints proposed for models of SWR were multiple access (all lexical items that are consistent with the input are activated), multiple assessment (the activated items are mapped onto the signal and top-down context in parallel), and real-time efficiency (i.e., a model should make optimal use of available information).

This last constraint is central. Rather than waiting for the best candidate to emerge by simple matching of phonemes to lexical representations, the model posits active removal of words from the "recognition cohort" (the set of activated candidates). Thus, as a word like *beaker* is heard, initially all words beginning with /b/ would be activated. When /i/ is heard, all items beginning /bi/ (*beaker, beetle, bead*, etc.) remain in the cohort, but words that mismatch (*baker*, *batch*, etc.) are removed. In the original model, a top-down mismatch (incompatibilities between the syntactic or semantic properties of the word and sentential context) could also remove an item from the cohort, making "Cohort I" an *interactive* model; although the model assumed bottom-up priority (top-down knowledge did not prevent items from entering the word-initial cohort, it only helped remove them), bottom-up processing was constrained directly by top-down knowledge.

These principles combine to predict that words will often be recognized prior to word offset: assuming clear speech as input, a word will be recognized prior to its offset if there is a unique completion prior to word offset, or if context provides sufficient listener confidence in the as-yet incomplete word. A key innovation in the Cohort model was its implicit segmentation strategy. Utterance onset marks the onset of the first word in a series. As one word is recognized, its offset marks the onset of the next item. The basic principles of the Cohort model, and in particular, the notion that segmentation would emerge from continuous mapping of phonemes to words, have motivated a tremendous amount of research and insight into SWR, and paved the way for subsequent models.

The model was revised slightly (Marslen-Wilson, 1987, 1989); "Cohort II" assumes selection must be *autonomous* from integration. This repairs problems with some predictions of Cohort I (e.g., predicting great difficulty recognizing words with low-probability relative to a context, such as, *I put on my hiking <u>beetle</u>*). The grain of input was increased from phonemic to

featural, to allow for a small degree of mismatch tolerance (about one feature), and activation was predicted to be related to goodness of fit weighted by word frequency.

We will turn now to simulation models, which have largely followed from the empirical findings of Marslen-Wilson and colleagues, and the processing principles articulated in the Cohort model.

2.3. Simulation models

Mathematical and verbal models can generate specific predictions when their underlying assumptions can be combined in a straightforward way (e.g., when stages of processing are clearly ordered and information only flows forward), especially if they do not address the fine-grained time course of lexical activation. When processing steps cannot be easily ordered or are expected to interact, or fine-grained time course predictions are desired, verbal models become unwieldy, and a mathematical model may be intractable or simply very difficult to derive analytically. In such cases, *simulations* with an implemented processing model (such as a neural network or production system) may be needed.

Simulation presents advantages but also challenges. While all models make simplifying assumptions, implementing a model requires explicit choices about inputs, outputs, and details that may not be part of any underlying theory, but are needed to make a simulating model work. Grappling with such details in order to create a simulation model may identify incorrect or incompatible assumptions that appeared reasonable in a verbal or mathematical model, or may reveal that aspects of human behavior emerge in unanticipated ways from the model. In this section, we review a handful of simulation models chosen to illustrate important developments in SWR modeling. Specifically, we review two "*hand-wired*" and four learning models. Parameters in "hand-wired" models are set by a researcher on the basis of (e.g., phonetic) principle, intuition, trial-and-error, or algorithmic search. More important than where the parameters come from is the fact that they are fixed by the modeler for a given simulation rather than learned.

2.3.1. Hand-wired models

2.3.1.1. TRACE (McClelland & Elman, 1986[iii]) was the first major implemented processing model of speech perception and SWR. It remains one of only a few *realist* (Kello & Plaut, 2003) models of SWR (see also Klatt, 1979, and Plaut & Kello, 1999, discussed below), and has by far the greatest depth and breadth of empirical coverage. TRACE extended the connectionist interactive activation framework (McClelland & Rumelhart, 1981) from reading to speech and was explicitly motivated by a desire to build and improve upon Cohort (McClelland & Elman, 1986, pp. 52-53). The model has three layers of units: featural, phonemic, and lexical (see schematic in Figure 1). Feature nodes are activated by input that roughly represents acoustic-phonetic properties of speech sounds by using 9 acoustic-phonetic feature continua, each represented by a bank of 7 units. Phoneme patterns are spread out over time, with features ramping on and off over 11 time steps (each corresponding to about 10 msecs). Because phonemes spread over many steps, but phoneme centers are only 6 steps apart, the input includes a coarse analog of coarticulation: on either side of a phoneme center, information about the current phoneme is added to that for the preceding or following segment, making the pattern for each phoneme context-dependent.

Feature nodes send activation forward to the phoneme layer, which consists of banks of phoneme templates aligned at multiple time slices (see more detailed schematic in Figure 2). This reduplication of units allows TRACE to handle the temporal extent of speech input by

spatializing time. Phoneme templates are maximally activated by a specific feature pattern aligned with them in time. Temporally overlapping phoneme units compete by lateral inhibition, such that ambiguous inputs will partially activate multiple phoneme units. However, competition will generally lead to a clear "winner" for each phoneme in the input (i.e., a phoneme unit that is substantially more activated than any others for "its" stretch of time).

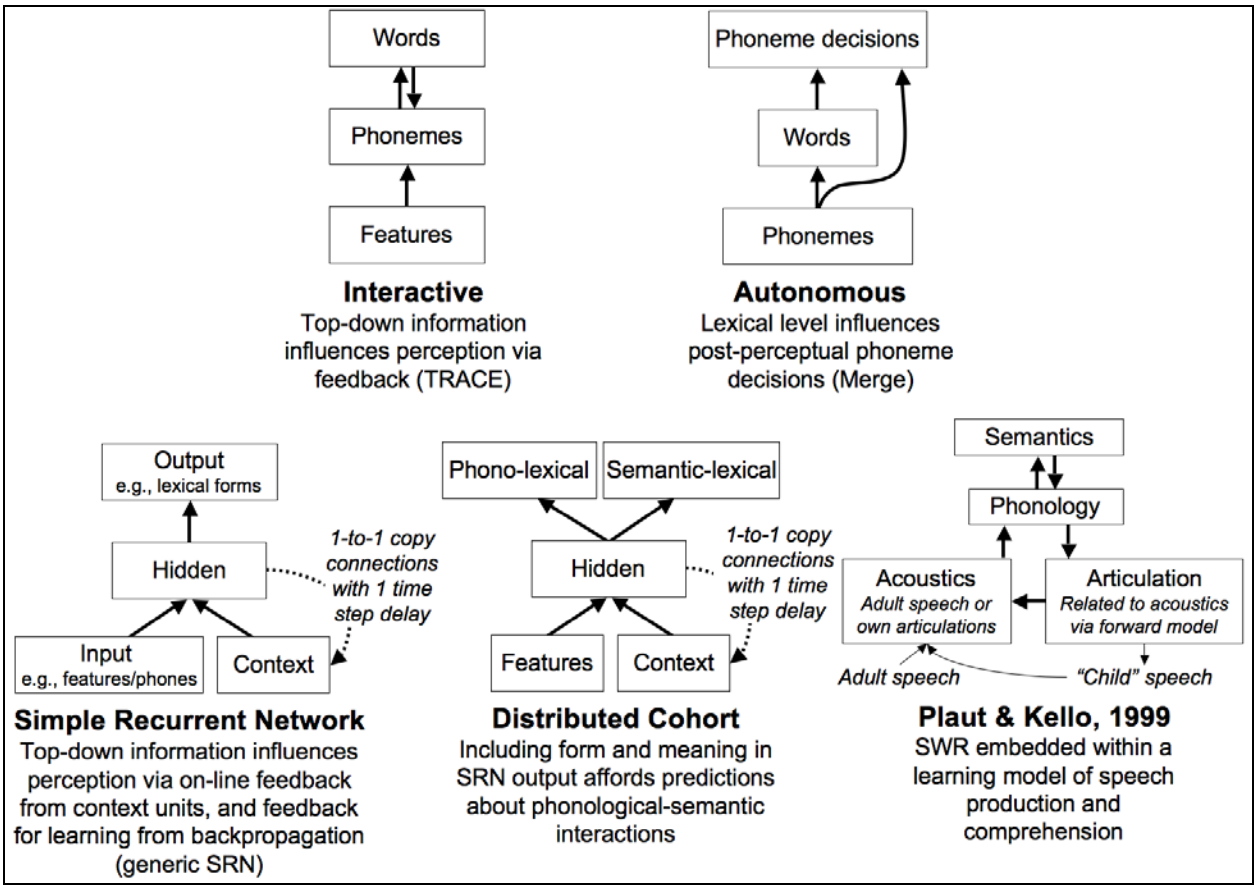

Figure 1: Schematics of five of the model types reviewed in this section. TRACE and Merge use localist representations; Distributed Cohort and Plaut & Kello use distributed representations; SRNs can use either.

The same scheme connects phonemes to words. Lexical templates are duplicated across time and are maximally activated when properly ordered phoneme units aligned with the template are maximally activated. Lexical units also compete with each other through lateral inhibition, with incomplete or ambiguous phoneme sequences partially activating multiple word units, and competition resolving ambiguity. A crucial feature of TRACE's architecture is feedback connections from lexical units to their constituent phoneme units (phoneme-to-feature feedback is typically disabled to speed processing, McClelland & Elman, 1986, p. 23). This feedback makes TRACE interactive (higher levels influence their own sources of input) and is one of the most controversial aspects of the TRACE model (discussed below; for recent debate see McClelland, Mirman, & Holt, 2006; McQueen, Norris, & Cutler, 2006; Mirman, McClelland, & Holt, 2006a).

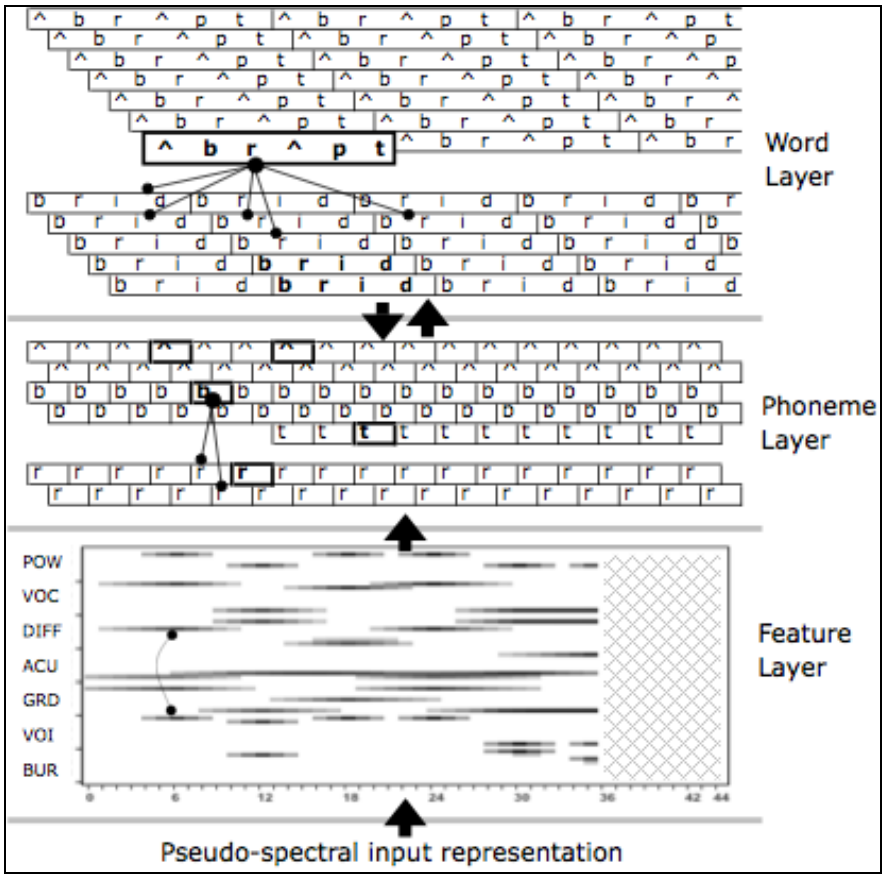

Figure 2: More detailed schematic of the TRACE model (adapted from Strauss et al., 2007) only showing four phonemes and two words. TRACE solves alignment and segmentation problems by reduplicating each word and phoneme node at multiple temporal alignments. Arrows stand for forward and backward connectivity (note the absence of phoneme-feature feedback, which is off by default in the model, but can be turned on). Nodes at low levels feed forward to larger units that contain them (e.g., featural patterns corresponding to voicing activate voiced phonemes, such as /b/; /b/ feeds forward to words that contain /b/), and nodes at higher levels feedback to the nodes from which they receive feedforward activation. Connections indicated with filled circles are inhibitory; nodes can inhibit other nodes at their own level ("lateral inhibition") if they overlap with them temporally.

TRACE differs from Cohort in that it eschews explicit consideration of mismatch or word boundaries, though it is implicitly sensitive to both. Activation in TRACE is based on *continuous mapping* of bottom-up matches to lexical representations. A bottom-up match to a lexical representation will send activation to that word even if there was an earlier mismatch (Allopenna, Magnuson, & Tanenhaus, 1998, capture this distinction with the terms *alignment* and *continuous mapping* models, where alignment models, such as Cohort, explicitly code mismatches relative to word onset). However, lateral inhibition makes the system sensitive to mismatches and, implicitly, to the position of the mismatch and details of the competition neighborhood. For example, an early mismatch is more penalizing than a late mismatch (given *candle* as input, nodes for *candy* or even *camera* or *cabin* will be activated more strongly than *handle*). This is because by the time the input overlaps with, e.g., a rhyme, items overlapping at

onset are already activated, and the rhyme must overcome lateral inhibition from the target and its onset cohort. Thus, the competitor set predicted by TRACE is intermediate between Cohort's and NAM's: onset overlap is an advantage, but items with initial mismatch may still be activated (an effect that is increased if there is uncertainty/noise in the input). Allopenna et al. (1998) found close fits between TRACE's predictions and the time course of phonological competition in human SWR (see Section 3.3).

TRACE depends on a fairly large set of parameters, such as the strength of bottom-up and top-down connections. Unlike most simulation models, where free parameters are fit to data, the TRACE parameters were fixed by McClelland and Elman, and have been used since then with only small changes. In the original paper, TRACE accounts for more than a dozen aspects of human speech perception and SWR, including categorical perception, segmentation of fluent, multi-word utterances, and lexical and phonotactic effects on phoneme recognition. Recent work has shown that TRACE also provides an excellent model of the fine-grained time course details of SWR (Allopenna et al., 1998; Dahan, Magnuson, & Tanenhaus, 2001; Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Spivey, Grosjean, & Knoblich, 2005). McClelland (1991) made an important refinement to TRACE – adding intrinsic noise – that allowed it to account properly for joint effects of context and stimulus (see Section 4.2).

Two aspects of TRACE have fueled the development of alternative models. The first is that the strategy of reduplicating phoneme and word templates to solve the temporal extent problem is arguably inelegant and implausible (cf. McClelland & Elman, 1986, p. 77). The second is the theoretical assumption of interaction (lexical-sublexical feedback, which we discuss in detail in Section 4).

2.3.1.2. Shortlist/Merge. Shortlist (Norris, 1994; Norris, McQueen, & Cutler, 1995) is a fundamentalist simulation model that combines aspects of autonomous, feedforward models like Race (Cutler & Norris, 1979) and Cohort II with the competition dynamics of TRACE. A primary motivation in the development of this model was to keep positive characteristics of TRACE (e.g., competition dynamics) while avoiding weaknesses (e.g., the large number of nodes and connections due to reduplication of nodes over time). In the first stage of processing, bottom-up activation generates word candidates aligned with each phonemic step of input (the bottom-up activation was originally intended to be from a simple recurrent network (SRN); in practice, a dictionary lookup is used). The best candidates (up to 30) at *each* phonemic input step form the *shortlist* at that position. The items from all shortlists are wired together into an interactive-activation competition network as each new phoneme is heard, and items that overlap in time inhibit one another (see Figure 3).[iv]

Shortlists are determined by match scores. Words get one point for every phonemic match, and -3 for every mismatch. To enter a shortlist, a word's score must be among the top 30 at a particular position. The mismatch penalty is so strong that the metric functions much like an alignment metric, allowing primarily onset-overlapping words into the shortlists. For example, when the input is *cat*, words beginning with /k/ are candidates at the first phoneme position when the first phoneme has been presented. When the second phoneme is presented, the shortlist at the first phoneme position is narrowed to words beginning /kae/, and words beginning with /ae/ are candidates at the second phoneme position. At the third phoneme, words beginning /kaet/ are candidates for the first phoneme shortlist, words beginning /aet/ are candidates for the second phoneme shortlist, and words beginning /t/ are candidates for the third phoneme shortlist. For words that rhyme or otherwise mismatch the input to enter the competitor set, the competition

neighborhood must be sparse and the input word must be long. That is, for an initial mismatch to be overcome, a rhyming word would have to match at the next four positions to arrive at a positive score and have some chance of entering the shortlist; e.g., given /kaet^lcg/ [*catalog*], /baet^l/ [*battle*] could enter the first-phoneme shortlist after /l/ is presented (assuming there were not already 30 words in that shortlist with match scores greater than one). A unique and crucial feature of Shortlist is the use of stress to constrain entry into shortlists (Norris et al., 1995; Norris, McQueen, Cutler, & Butterfield, 1997). This feature could (and should) be added to other models.

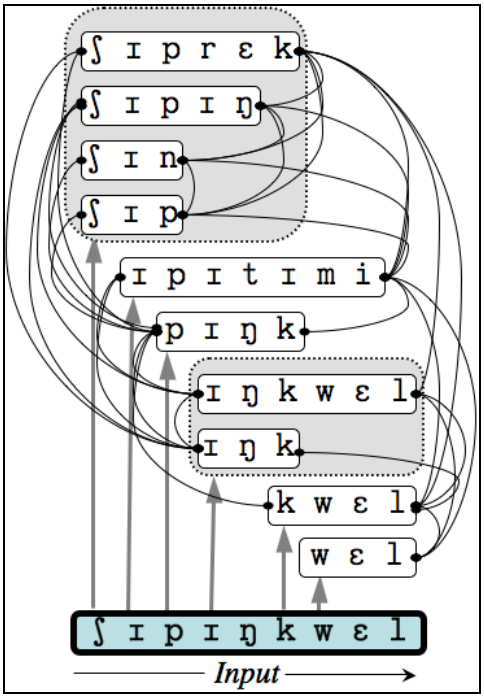

Figure 3: Lexical competition in Shortlist. The shaded box at the bottom shows the input ("ship inkwell"). As the input is presented to the model, shortlists of items with positive match scores are constructed for each phoneme position (up to a maximum of 30 items per position). Arrows indicate aligned shortlists. For most positions in this figure, only a single item from the shortlist is shown. Larger subsets of the complete shortlists are shown at positions 1 and 4 (shaded groups). Items compete with other items that overlap with them at any position – including items in other shortlists. All inhibitory connections are shown. Only word pairs that do not overlap temporally do not have inhibitory connections. The shortlists shown are an idealization of what might be active at word offset, but are not taken directly from a simulation.

This division of labor between lexical search and competition allows Shortlist to use many fewer connections than TRACE. Shortlist is sometimes claimed to require fewer nodes than TRACE as well, but this depends on the nature of the lexical search mechanism. If an SRN were used, the entire lexical search network would have to be replicated at each input step – since a new lexical search is generated for *every* input position as each phoneme is presented. This would result in at least the same number of lexical representations as in TRACE. However, since SRNs also predict a variety of lexical competition effects (Magnuson, Tanenhaus, & Aslin, 2000), there would appear to be no need either for multiple SRNs aligned with each phoneme,

nor for an interactive activation network – a single SRN would simultaneously provide lexical search and competition.[v]

Shortlist is a fundamentalist implementation of the theoretical principle that word recognition can be achieved efficiently with a modular division of labor between initial access and selection via competition. It is a fundamentalist model because it incorporates only details necessary for testing those primary assumptions. The **Merge model** (Norris, McQueen, & Cutler, 2000) is a separate but related fundamentalist model that is also purely feedforward. Merge is consistent with the Shortlist framework, but was designed to examine whether lexical effects on phoneme decisions can be predicted without lexical-phonemic feedback, by adding post-lexical phoneme decision nodes (see Figure 1). Merge is meant to be roughly equivalent to the competition network of Shortlist, though it is greatly simplified. Merge has only been demonstrated with a few phonemes and words – up to 4 words and around 6 phonemes, depending on the simulation. While the input to the model has a subphonemic grain – phonemes ramp on over three time slices – the architecture does not encode temporal order. Word units have undifferentiated connections from phonemes, such that the inputs *dog*, *god, odg, ogd, dgo,* and *gdo* would all activate lexical units for *dog* or *god* equally well. All the same, the model qualitatively accounts for several results that previously had been thought to require interaction (see Section 4).

New, Bayesian versions of Shortlist and Merge have been proposed (Shortlist B and Merge B; Norris & McQueen, 2008). Shortlist B resides at Marr's (1982) computational level of information processing theories, providing a description of a putatively optimal mapping from speech input to spoken words. It has an unusually fine grain for a computational level theory: diphone confusion probabilities from a gating task are used to construct phoneme likelihoods at a subsegmental grain. Those likelihoods are conditioned on lexical knowledge and potentially other context, but with the stipulation that the mechanism for combining these information sources must operate without feedback. This is not a stipulation commonly found in Bayesian approaches to perception; for example, Rao (2004, 2005) demonstrates how Bayesian inference can be implemented in a neural architecture employing feedback, affording optimal combination of top-down and bottom-up information sources, and close fits to behavioral data. Movellan & McClelland (2001) have also proven that the interactive activation framework of a model like TRACE can implement an optimal Bayesian process. Nonetheless, the approach taken with Shortlist B has the potential to generate extremely precise predictions and may lead the way to new approaches. We return to the controversial question of whether feedback occurs in speech perception and SWR in Section 4.

2.3.2. Learning models

2.3.2.1. Simple recurrent networks. SRNs (Elman, 1990) have been applied to SWR with limited coverage. A basic SRN consists of four sets of units: input, hidden, output and context (see Figure 1). There are feedforward connections from input to hidden units and from hidden to output units, as in a standard feedforward network. The context units contain an exact copy of the hidden units at the previous time step and are fully connected to the hidden units (or, equivalently, each hidden unit has a recurrent connection to all other hidden units with a delay of 1 cycle). This innovation of recurrence, or feedback, provides the network with a limited potential memory for previous time steps. All of the connections (except hidden-context, which are one-to-one copy connections) are trained via backpropagation (where actual input is compared to observed output, and connections receive "blame" for the discrepancy based on how

much of the error they contributed, and their weights are changed proportionally; Rumelhart, Hinton, & Williams, 1986). A typical approach is to present a sequence of input vectors corresponding to a sequence of single phonemes and set the desired output to be the next phoneme or the current word. Depending on the nature of the training set and the size of the network, SRNs can develop sensitivity to fairly long stretches of context. While a common approach is to use a series of phonemes as input and localist lexical nodes as output, one can of course use distributed representations, or change the task to predicting the next phoneme, or even the previous, current, and next phonemes. These choices have a significant impact on what the model learns.

Norris (1990) reported SRN simulations in which words in a small lexicon that overlapped at onset activated each other, but words that mismatched at onset and overlapped at offset did not, consistent with predictions of the Cohort model. Although this is a logical result, given the model has explicit access to ordered input, Magnuson et al. (2000) showed that it depends on the training regimen. If the model is given perfectly clear inputs and is trained until error rate asymptotes (the procedure followed by Norris), it will only show onset competition. If instead training continues only until every word in the lexicon is "recognized" correctly using a simple, minimal threshold (only about one-fifth as much training), the network exhibits rhyme effects, and will also more easily learn new words and be more tolerant of noisy inputs. Furthermore, early in training, the model shows roughly equivalent rhyme and cohort competition; adults learning novel neighborhoods of words show the same progression (Magnuson, Tanenhaus, Aslin, & Dahan, 2003).

There is disagreement about the nature of the architecture of SRNs. Some claim that SRNs are not interactive (Cairns, Shillock, Chater, & Levy, 1995; Norris, 1990), since the input units are not influenced by the output level. Others disagree (e.g., Magnuson, McMurray, Tanenhaus, & Aslin, 2003a; McClelland et al., 2006) on the basis that recurrent connections allow context to have a direct influence on the earliest stage of processing (since feedback from context is mixed with bottom up input at the hidden unit level), even if the mechanism does not include feedback from explicitly *lexical* nodes. Specifically, the input to the hidden layer at each time step is the current bottom-up input *and* an exact copy of the hidden unit states from the previous time step; the latter are a the result of multiplying the previous input and context by the hidden unit weights, so the input includes the output of the first of the two feedforward transformations the model performs.

In summary, SRNs avoid problems of TRACE (reduplicated units, inability to learn), and have the potential to be the basis of a "next generation" of models. Indeed, the next two models are based on this architecture.

2.3.2.2. Distributed Cohort Model (DCM). Gaskell and Marslen-Wilson (1997) began pushing beyond the typical focus on sound form recognition by incorporating simultaneous semantic representations in their model. The input (binary phonetic features), hidden, and context layers followed standard SRN design. Their innovation was the use of two output layers: "phonology" (phonological form) and "lexical semantics" (an arbitrary, sparse binary vector; see Figure 1). Gaskell and Marslen-Wilson (1999) argued that distributed representations and simultaneous activation of phonological and semantic dimensions of words provide fundamentally different ways of thinking about competition. In localist models such as TRACE, when the input supports two lexical items, there is explicit activation of both representations (different nodes at the lexical layer) and explicit competition between them (through mutually

inhibitory connections between the lexical units). In a distributed model, all items are represented with the same set of nodes; thus, both activation of and competition between multiple representations is implicit in the *blend* formed by the competing patterns.

Gaskell and Marslen Wilson (2002) tested a prediction that follows from this conceptualization. Given a word fragment with semantically unrelated phonological completions (e.g., /kaept/ can begin *captive* or *captain*), the system can settle on a *phonological* pattern, but semantic activations will be a blend of the semantics for the phonological competitors. Thus, such a fragment should produce phonological (repetition) priming, but not semantic priming. In contrast, if few completions are possible (e.g., /garm/ can only be *garment*), the system will settle on single phonological and semantic patterns, and both phonological and semantic priming should be observed. This is precisely what Gaskell and Marslen-Wilson found.

Gaskell and Marlsen-Wilson (2002) claimed that only a distributed model could account for such differential activation, though it appears DCM does so by virtue of including both phonological and semantic outputs, not by virtue of using distributed representations. If semantic representations were added to TRACE (e.g., if the phoneme layer simultaneously fed to the current lexical [phonological form] layer, and to a layer of semantic primitives that fed forward to a second lexical [semantic form] layer), it would make similar predictions: /kaept/ would activate mutually reinforcing units (*captain* and *captive*) on the phonological side, predicting strong phonological priming, but /kaept/ would activate disparate semantic representations, and predict weak semantic priming. Although localist and distributed models may not make conflicting predictions for currently known empirical results, there are strong arguments for preferring distributed to localist representations (Masson, 1995; Plaut, McClelland, Seidenberg, & Patterson, 1996) and the DCM represents a crucial step in that direction among SWR models.

2.3.2.3. PK99. Plaut and Kello's (1999) model is perhaps the most ambitious model of SWR yet proposed, and it is embedded within a comprehensive model of the development of speech production and speech comprehension (see Figure 1). The model learns to control a set of articulatory parameters to generate acoustics based on "adult" input (well-formed acoustics) and self-input (acoustic results of its own articulations). The acoustics are fairly close analogs of the speech signal (formant frequencies and transitions, frication, plosiveness, loudness, and the visual feature of jaw openness). The model learns a bi-directional mapping between acoustics and articulations and the mapping from both of these phonological representations to an arbitrary set of semantic patterns. The first report was extremely promising; in the domains tested, the model exhibited a range of desirable learning and processing behaviors. We hope development of this model continues, as we find that it provides the most promise for significant progress in modeling the development of speech production and comprehension.

2.3.2.4. Adaptive Resonance Theory (ART). ART is a powerful connectionist learning framework. Inputs are initially mapped to early representations in a working memory stage. These then map (through bidirectional links, allowing feedback) to *list chunks* (combinations of lower-level units that have co-occurred over learning). Chunks of equal length inhibit each other and longer chunks "mask" smaller chunks that are contained within them. The framework has allowed for an impressive array of fundamentalist models (separate models for processing aspects of real speech [ARTSTREAM; Grossberg, Govindarajan, Wyse, & Cohen, 2004], phonological patterns [ARTPHONE, Grossberg, Boardman, & Cohen, 1997], and word segmentation [ARTWORD, Grossberg & Myers, 2000]), which suggests great promise for a

comprehensive, realist model, but such a model has not yet been reported (see also Goldinger & Azuma, 2003, for suggestions of how Goldinger's [1988] episodic lexicon model might be combined with the ART framework).

An intriguing aspect of ART's processing assumptions is that its version of top-down feedback cannot cause hallucinatory representations. A "2/3 rule" means that weak inputs (e.g., phonetic features corrupted by noise) can be strengthened once recognized by higher levels of processing, but completely absent inputs cannot be created from nothing. As we discuss below, a common criticism of feedback in TRACE is that it could make the system hallucinate (Norris et al., 2000). Although, in practice, misperception in TRACE seems generally similar to misperception in humans (Mirman, McClelland, & Holt, 2005) and the default TRACE parameters also give it strong, bottom-up priority, future modeling efforts might benefit from nonsymmetrical feedback rules such as those implemented in ART**.**

3. Evaluating and comparing models

The recent history of SWR includes disagreements about whether particular models succeed or fail to account for various phenomena. There has been a salient absence of agreed upon principles for gauging model success or failure and for comparing models. We will argue that assessing success requires (1) clear *linking hypotheses* (links between the tasks performed by human subjects and the measurable properties of a model), and (2) attributing a success or failure to one of four levels (in decreasing order of importance): theory, implementation, parameters, or linking hypotheses. After introducing these issues, we will illustrate them with recent examples from the literature, and propose a set of candidate principles for assessing success and comparing models. These principles will frame a larger discussion of the feedback debate in Section 4.

3.1. Linking hypotheses.

The first question for comparing model behavior to human behavior is how to link properties of the model to the task performed by human subjects. The simplest approach is to look for qualitative similarity between a model and human data. For example, if lexical node activations correlate inversely with human response times and error rates in some task, it is reasonable to accept this as a model success, though this is a weak standard. One would do better to ask whether the model also provides good *quantitative* fits, and whether the fits are to condition means or individual items (e.g., does it predict errors on the correct class of items, or depending on the task, does it predict appropriate errors?). As the quantitative fit and grain of prediction increases, so should the standard for success. The standard can be strengthened further by examining how closely the model's task resembles the human subjects' task by establishing explicit linking hypotheses: concrete operational definitions tying features of model performance to human behaviors and tasks.

Linking hypotheses typically receive little attention. However, one cannot say a *model* has failed unless one has first appropriately linked (a) model performance to human performance, (b) stimulus materials for human subjects to model materials, and (c) task constraints faced by humans to task constraints on models (e.g., through choice models).

3.2. Model successes and failures: Levels of analysis.

A model success or failure can be linked to one of four levels of decreasing importance: theory, implementation, parameters, or linking hypotheses. As we have just discussed, a "failure" or "success" due to improper linking hypotheses is not informative in the same way that an

experimental failure due to improper operational definitions is not informative. A failure at the level of theoretical assumptions is of greatest interest and holds the greatest possibility for progress (i.e., theory falsification). Before a model failure can be attributed to underlying theoretical assumptions, one must establish that the failure cannot be attributed to implementational details or to parameter settings. Implementational details include factors such as input representation, numbers of units in a neural network model, and details of processing dynamics (e.g., activation functions). Parameter settings play a critical role in simulating models, so, for example, if TRACE model simulations suggest competitors are inhibited too much, one cannot conclude that lateral inhibition is fundamentally flawed without testing different values of lexical inhibition, phoneme inhibition, bottom-up excitation, etc. Likewise, in a learning model, performance may change radically as a function of amount of training, as we mentioned above in our discussion of SRNs.

Parameters are of particular importance, as there have been suggestions that a model as complex as TRACE should only be tested with minor deviations from the original parameter set. It is true that if different parameter sets are used to model different results, the model loses its generality -- the breadth of model successes cannot be attributed to underlying theoretical assumptions if each success requires different parameters. On the other hand, equating a model with a parameter set produces a similar problem: the model loses generality because the constraints of the parameter set are placed on a par with underlying theoretical assumptions. The simple alternative is not to limit model explorations to a "standard" parameter set, but the onus is on the modeler to test whether parameter changes needed for one phenomenon prevent the model from fitting results it was known to fit with the previous settings.[vi]

We will now review a case where proper linking hypotheses provide insight into how task constraints shape behavior, and other cases where apparent model failures were actually due to improper linking hypotheses. Then we will turn to candidate principles for gauging success and comparing two models.

3.3. Improving models with linking hypotheses.

An interesting outcome of the use of simulation models is that for more than a decade, models made predictions at a finer grain than could be tested with standard psycholinguistic tasks. Models like TRACE (McClelland & Elman, 1986) make explicit predictions about the parallel activation of similar items and the time course of competition between them. For example, panel B of Figure 4 shows the complex pattern of activation and competition among TRACE's lexical nodes for items like *beaker, beetle, speaker,* and *carriage* when the input is an item like *beaker*.

Fine-grained lexical activation predictions began to be testable with the advent of the "visual world" eye-tracking paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In this paradigm, participants see multiple objects, and their eye movements are tracked as they follow spoken instructions to perform visually guided movements (e.g., "click on the beaker"). At any instant, participants can fixate only one object, but time course can be estimated from average fixation proportions over time. Panel A of Figure 4 shows data from Allopenna et al. (1998), who presented subjects with displays of four items like (on critical trials) *beaker, beetle, speaker,* and *carriage*, and examined fixations as subjects followed an instruction like *click on the beaker*. While there is an obviously strong qualitative fit between the data and the TRACE activations in panel B, Allopenna et al. established a closer link by linking model time to real time (by relating average phoneme duration in real speech materials to TRACE cycles per

phoneme) and, more importantly, by explicitly considering task constraints on human subjects (panel C). Subjects had four possible fixation outlets – the pictures on the screen. Allopenna et al. assumed that bottom-up lexical activation was not restricted to the displayed items, and based lexical activation on activation and competition in the entire TRACE lexicon. To incorporate the four-choice task constraint, they computed response probabilities based only on the activations of the four displayed items (using a variant of the Luce [1959] choice rule). With one free parameter (a multiplier used in the choice rule[vii]), this linking hypothesis greatly improves fit – by taking into account task constraints faced by human subjects, and simultaneously providing a "placeholder" model of the decision process (in the sense that it is obviously incomplete). It also suggests the possibility that TRACE activations may surprisingly closely approximate human lexical activations, as a very simple linking hypothesis taking task constraints into account results in high model-data fits (and this same linking hypothesis allows close fits of changes in looking behavior when cohort competitors are present or absent; Dahan, Magnuson, Tanenhaus, & Hogan, 2001).

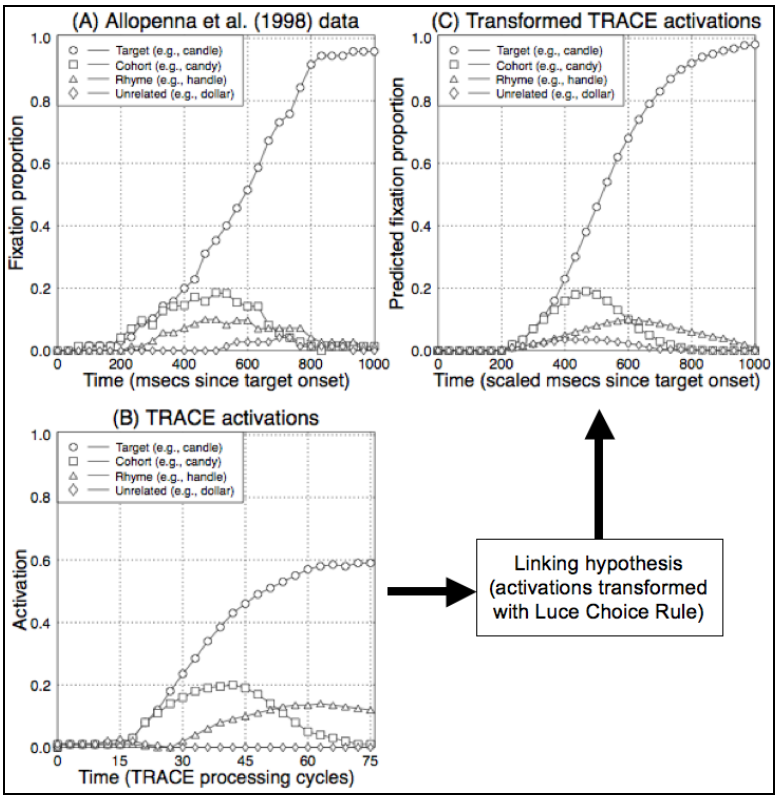

Figure 4: Comparison of eye tracking data (A), TRACE activations (B), and TRACE activations transformed into predicted response probabilities via explicit linking hypotheses (C). Adapted from Allopenna et al. (1998).

Allopenna et al. calculated fit with r^2 (do human and model proportions rise and fall together?) and root mean squared (RMS) error (are the actual values close?); r^2 was high and RMS was low.[viii]The Allopenna et al. (1998) study provided partial resolution to a paradox having to do with similarity metrics (see Figure 5). In tasks like cross-modal semantic priming (e.g., Marslen-Wilson, 1990), there is strong evidence for onset (or "cohort") competition (e.g.,

hearing *beaker* primes *insect*, an associate of *beetle*, as *beetle* is strongly activated by phonological similarity to *beaker* and then activates insect via spreading semantic activation), but not for rhyme competition (*beaker* would not detectably prime *stereo*, an associate of *speaker*). In contrast, NAM's similarity metric includes rhymes, and NAM provides the best available predictions for large sets of items (accounting for about 15% of the variance in SWR tasks). TRACE makes an intermediate prediction: onset competitors have an advantage because they receive substantial bottom-up activation without strong inhibition during the early part of the word. Rhymes are predicted to be activated, but to be at a significant disadvantage: by the time they have bottom-up support, the target and onset competitors are sending strong inhibition. Since the eye tracking data matches TRACE's predictions so closely, this suggests that rhymes are activated, but more weakly than onset competitors. In cross-modal semantic priming, effects depend on phonologically-based activation spreading semantic activation. If rhyme activation is weak, it is not surprising that it is difficult to detect it in a mediated task. This case illustrates the symbiotic role of models; this level of resolution of the paradox could only have been attained by use of both quantitative empirical methods and an implemented model with an intervening linking hypothesis.

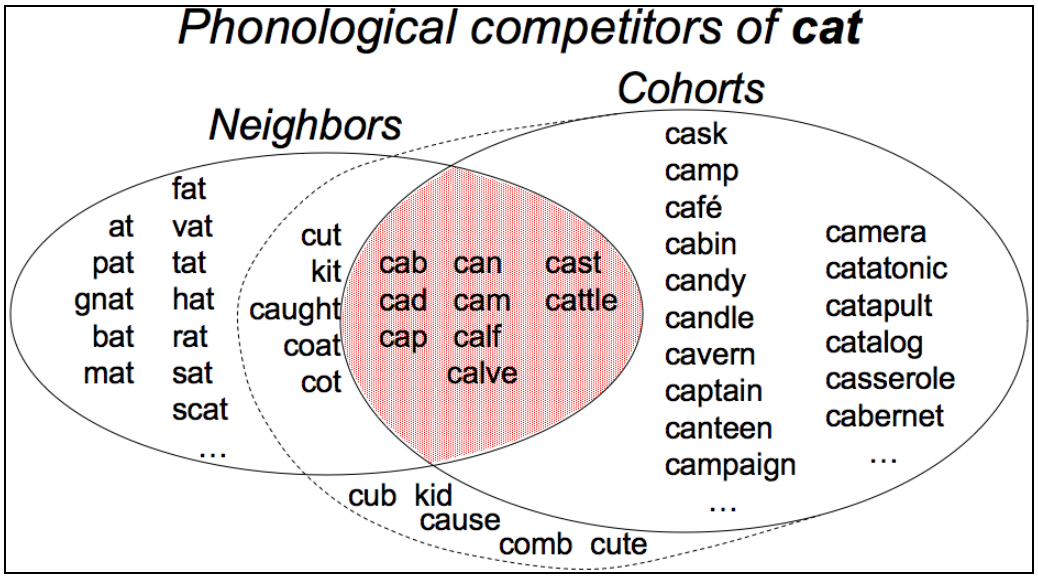

Figure 5: The relationships of similarity metrics. Neighbors differ from each other by a single phoneme. Cohorts overlap at onset. Often, the overlap threshold is 200 msecs or approximately the first two phonemes. Less often, overlap in the first phoneme is the threshold (delineated by the dotted curve). The shaded region indicates items that are both neighbors and cohorts. TRACE predicts strongest activation for items that are both (2-phoneme overlap) cohorts and neighbors, then for (2-phoneme overlap) cohorts, then other neighbors, and little activation of items overlapping in a single onset phoneme (though greater activation is predicted for items like *cut* with a single mismatch versus *cub* with two mismatches).

3.4. Linking to human materials and task constraints.

Marslen-Wilson and Warren (1994) examined the role of lateral inhibition as a competition mechanism in TRACE by creating cross-spliced versions of a word like *net* that combined the initial CV of one word and the final C of another (e.g., the initial CV of *neck* plus

the final C of *net*) or the initial CV of a nonword (*nep)* and the final C of the word (*net*). These cross-spliced items included misleading coarticulatory (subcategorical) information about the final C. The baseline item was the initial CV of the target spliced onto the final C of another recording of the same word. The three conditions were labeled W1W1 (CV of one recording of *net* spliced onto the final C of another recording of *net*), W2W1 (*neck* + *net*), and N3W1 (*nep* + *net*). Marslen-Wilson and Warren's simulations indicated that TRACE predicted the following response time pattern: W1W1 < N3W1 < W2W1 (with a large increase for W2W1), but human lexical decision data showed the pattern W1W1 < N3W1 ≈ W2W1. Marslen-Wilson and Warren attributed this discrepancy to lateral inhibition in TRACE, which they argued was too strong. Norris et al. (2000) ran simulations with Merge and a radically simplified interactive activation model (the 6-phoneme and 4-word Merge model with lexical feedback). Merge successfully predicted the response time pattern, as did their interactive analog, but only if it was made to cycle multiple times at each input step, effectively increasing the amount of inhibition that occurred prior to a decision. So Marslen-Wilson and Warren argued TRACE had too much competition, while Norris et al. argued that competition in TRACE was too slow.

Dahan, Magnuson, Tanenhaus, and Hogan (2001) revisited this paradigm with eye tracking paired with TRACE simulations (see Magnuson, Dahan, & Tanenhaus, 2001, for more simulation details). Contrary to the lexical decision data, they found that fixation trajectories fit the pattern W1W1 < N3W1 < W2W1 (though the pattern was not as extreme as in the Marslen-Wilson & Warren simulations). Magnuson et al. explained the discrepancy between the eye movement and lexical decision data by assuming a "yes" response could be triggered if the activation of *either* W1 or W2 reached a threshold. This would decrease average response time for W2W1, assuming the activation of W2 (*neck*) generates infrequent "yes" decisions. In separate lexical decision simulations based on eye movement time course and TRACE activations, there were ranges of parameters where this simple assumption leads to correct RT predictions (W1W1 < N3W1 ≈ W2W1). Contrary to the Marslen-Wilson and Warren and Norris et al. simulations, new TRACE simulations correctly predicted the data at a very fine grain. Dahan et al. explained the discrepancy between their TRACE simulations and Marslen-Wilson and Warren's by deducing that the latter cross-spliced the TRACE stimuli much too late. Dahan et al. cross-spliced at the latest position possible that still led to the correct recognition of the intended final target. When the splicing is done as late as that reported by Marslen-Wilson and Warren, W2 is recognized rather than W1 given W2W1, and W2 is also recognized in a nonword condition (W2N1). If this happened with human subjects, the materials would be scrapped and replaced. This illustrates an important principle: the same care that is taken with materials for human subjects must be taken with model testing, in order to ensure adequate analogs between human and model conditions. The lexical decision simulations demonstrate that linking hypotheses can radically alter the apparent success or failure of a model.

3.5. Intuition and logic vs. simulation.

Consider the following predictions about TRACE and SWR in general. If word frequency has a pre-lexical locus, it should have a constant effect, detectable in both fast and slow word recognition responses. If frequency is a post-lexical decisional bias, frequency effects might disappear when subjects respond very quickly – before they hit the stage where frequency is integrated with lexical activation. Connine, Titone, and Wang (1993) found that indeed, frequency effects tend not to be detectable in fast responses and concluded that in a model like TRACE, such a result could only occur if frequency were a post-lexical bias. Dahan, Magnuson,

and Tanenhaus (2001) augmented TRACE with three frequency mechanisms: post-lexical (frequency applied in the choice rule rather than activations), resting levels (each word's activation in the absence of input was proportional to frequency), and bottom-up connection strengths (phoneme-word connections were proportional to word frequency). The intuitive expectation was that the latter two would lead to similar predictions, and both would differ from the first.

Dahan, Magnuson, & Tanenhaus (2001) compared time course predictions from TRACE to fine-grained time course measurements of frequency effects using eye tracking and the visual world paradigm. Empirically, human listeners showed a continuous influence of frequency that increased as more of a word was heard. Contrary to Connine et al.'s (1993) predictions, all three frequency-augmented versions of TRACE could fit the human fixation proportion data fairly well. Also surprisingly, the resting level and post-lexical mechanisms made virtually identical predictions, with a constant frequency influence. (To predict a late influence, the post-lexical account would require an additional parameter specifying when frequency should be applied.) The bottom-up connection weight mechanism predicted that the effect would be proportional to the amount of evidence, and provided the closest fit to the human data (especially the early time course). This mechanism would account for the Connine et al. results as a matter of task sensitivity: if you sample early in processing (with fast decisions) the magnitude of the frequency effect would be small. If the sensitivity of the task used were low (as it arguably is in lexical decision), a null result in early responses would not be surprising.

This example demonstrates the value of simulations with complex models over intuition-based expectations. Whenever possible, expectations should be verified with model simulations (see the Appendix for a list of tools that can be used for SWR simulations).

3.6. Comparing models.

Assuming two models account for overlapping phenomena, how should we compare them? First, if one appears to fail on some phenomena, the *level* of the failures must be identified, as we have just discussed. If the failures can be argued to be nontrivial, and all else is equal about the models, one has a basis for preferring the one with fewer failures. However, if all else is not equal (e.g., one model uses more realistic input or mechanisms, or one requires different parameter settings for different phenomena), one should prefer the model with greater realism, greater depth and breadth of coverage, or greater parameter stability.

A recent trend in model analysis has been to distinguish between models that fit empirical data because of inherent properties of the model from models that fit only because of specific parameter settings. The standard test of model performance is to compare model and human behavioral data under one specific set of parameter values (or a small range of values). However, a model may be flexible enough to fit any possible data. Ideally, model behavior should be fairly stable over parameter changes and the optimal parameter range should account for a relatively large set of behavioral data (i.e., parameter changes should not be required for each new behavioral data pattern).

Pitt and his colleagues have recently developed a method (called Parameter Space Partitioning, or PSP) for comparing models based on their performance across their parameter space (Pitt, Kim, Navarro, & Myung, 2006). PSP examines the range of qualitative data patterns (e.g., an ordering of RTs in different conditions) that a model is capable of producing and computes a partitioned map of parameter space in which each partition corresponds to a qualitatively different data pattern generated by the model. This allows one to assess whether a

good fit by the model is due to intrinsic properties that follow from the theoretical assumptions of the model, or merely to particular parameter settings. To conclude that a model is reasonably constrained (and cannot predict arbitrary data patterns), the following should hold: (a) the model should produce relatively few data patterns across the parameter space; (b) the empirically observed pattern (human data) should correspond to a relatively large proportion of the parameter space; and (c) most other data patterns the model can produce should be similar to the empirically observed data pattern, with relatively smooth changes in patterns from partition to partition (rather than radically different patterns).

Table 1: Candidate principles for evaluating and comparing models.

Heuristics for Evaluating Models

1. Model failures should not be accepted lightly
   a. If there is a qualitative failure, determine level of failure
      i. Theoretical (the underlying assumptions are wrong)
      ii. Implementation (an architectural or representational assumption is wrong)
      iii. Parameters (the model could fit the data with changes in parameters, but then previous model predictions must be verified with the new settings)
      iv. Linking hypotheses (are human and model materials and tasks comparable?)
   b. Failures of theory or implementation are strong evidence against a model
   c. Failures of parameters are strong evidence against a model only if new parameters are needed for each new data set
   d. Failures due to improper linking hypotheses are not model failures
2. In gauging degree of success, strong standards should be preferred to weak standards
   a. Quantitative fits are stronger than qualitative fits
   b. Item-specific predictions are stronger than condition-specific predictions
   c. Specific error predictions are stronger than error rate predictions
   d. Constrained models (based on parameter space partitioning) are stronger than unconstrained models (i.e., models that can fit patterns quite different from human performance)

Heuristics for Comparing Models

The heuristics cannot be strictly ordered; for example, disparity in heuristic (c) might outweigh heuristics (a) and (b)

In comparing two models:

a. Prefer the model with greater breadth (range of phenomena it models)
b. Prefer the model with greater depth (the model that can be held to a stronger standard of success, as in (2) above)
c. Prefer the model with greater realism (e.g., a model with more realistic inputs or outputs)
d. Prefer the more realistically constrained model (e.g., based on *parameter space partitioning*; see text)
e. When all else is equal, apply Occam's razor: prefer the simpler model

Parameter space partitioning offers a powerful tool for testing and comparing models. However, its results are only as good as the characterizations of models and problems it is given. For example, PSP is extremely computationally intensive, which limits the complexity of models to which it can be applied. When Pitt et al. set out to compare the TRACE and Merge models, they used a "toy" implementation of TRACE like that used by Norris et al. (2000) (with phonemic input and only a subset of TRACE's phonemes, a very small lexicon, and no ability to represent temporal order). This implementation might better be characterized as an extreme fundamentalist version of an interactive model, as it has little in common with TRACE aside from feedback. Similarly, they focused on tests of lexically-mediated phoneme inhibition (reviewed in section 4.2, below), but based the human behavioral standards on a report by Frauenfelder, Segui, and Dijkstra (1990), which has several problems (Mirman et al., 2005; see

Section 4.2), thus undermining their model analysis. Nonetheless, if candidate models are correctly implemented and human performance is correctly characterized, global qualitative model evaluation approaches such as PSP can offer important new insights into processes underlying SWR.[ix]

Conclusions. Currently, there are no generally agreed upon principles for evaluating individual models or comparing two models. Table 1 lists a candidate set of heuristics for model evaluation and comparison (Jacobs & Grainger, 1994, provide a more detailed set of principles). However, comparing two models is more difficult than one might expect, especially if they differ in realism and empirical coverage. To illustrate this, we will review a currently central debate in SWR, as an example of how model comparison takes place in the literature.

4. The feedback debate

Proponents of interaction in SWR (feedback connections from lexical to sublexical representations) argue that feedback (a) is a logical way to account for the many lexical effects on sublexical tasks that have been reported in SWR (for examples, see McClelland et al., 2006, and Mirman et al., 2006a), (b) makes a model robust to external or internal noise, and (c) provides an implicit representation of sublexical prior probability at multiple scales (e.g., biphone, triphone, … *n*-phone). Proponents of autonomous architectures – those with only feedforward connections – argue (a) feedback is unnecessary to account for lexical effects, (b) it cannot improve recognition, and worse, (c) feedback precludes truly veridical perception and predicts perceptual hallucination.

Proponents of the autonomous view have argued against feedback in two ways. First, they argued that all observed lexical effects on sublexical tasks can be explained by post-lexical integration of lexical and sublexical information (Norris et al., 2000). More recently, Norris and McQueen (2008) have argued that lexical and other contexts *should* influence word recognition under certain conditions, but only by means of a Bayesian decision process that has pre-perceptual access to context-conditioned probabilities (via an as-yet unspecified mechanism). What is required to falsify the autonomous position is empirical data showing lexical influence on pre-decisional sublexical processing. This has turned out to be a nontrivial enterprise in terms of developing experimental paradigms that proponents of both views would find convincing (for discussion see Dahan & Magnuson, 2006; McClelland et al., 2006; McQueen et al., 2006). Here, we will focus on model-specific issues that have been important in this debate.

4.1. What good can feedback do?

Norris et al. (2000; also Norris & McQueen, 2008) assert that feedback cannot possibly aid recognition. It can neither speed processing nor improve accuracy. Since there is no way to increase the information available in the signal, a system could not do better than simply activating the word with the best bottom-up fit to the signal. One piece of evidence they cite as support comes from TRACE simulations (Frauenfelder & Peeters, 1998; FP98) in which the usefulness of feedback was studied by comparing performance with feedback on and off. For the 21 words tested, about half were recognized more quickly with feedback, and about half were recognized more quickly without feedback. Thus, even in TRACE, the flagship interactive model, feedback seemed not to improve recognition.

Magnuson, Strauss, and Harris (2005) revisited this result, with three motivations. First, the general argument about the usefulness of feedback can be challenged on logical grounds (since, for example, words provide an implicit coding of prior probability for sublexical

phoneme sequences). Second, the FP98 simulations do not address a central motivation for feedback in interactive systems: feedback makes a system robust against internal or external noise. That is, feedback is useful because it affords context sensitivity, by implicitly coding prior probabilities of causes (phonemes, words, etc.), which can be especially useful given uncertain input. Given a sequence of phonemes including noise or ambiguity, the system could perform more quickly and/or accurately if it allowed context (lexical, syntactic, discourse, etc.) to help disambiguate the input as soon as possible. Third, the FP98 simulations only used a small set of words with particular properties (7 phonemes long, with a uniqueness point at the fourth segment). These were chosen for other simulations presented in the same chapter, but are not representative of the lexicon.

Magnuson et al. tested performance with and without feedback on a large (901-word) lexicon with several levels of noise added to the input. At every level of added noise, average accuracy and recognition time were better with feedback on. Without noise, nearly 75% of the lexical items were recognized more quickly with feedback on. Cases where words were recognized more quickly *without* feedback resulted from complex neighborhood characteristics; however, when noise was added, feedback preserved accuracy even for these items.

Another benefit of feedback is that it allows top-down knowledge to guide tuning or recalibration of the perceptual system when there are systematic changes in the input; for example, adjusting to a speaker with an unfamiliar accent. There is strong behavioral evidence that listeners use lexical information to tune the mapping from auditory to phonemic representations (Norris et al., 2003; Kraljic & Samuel, 2005, 2006; McQueen, Cutler, & Norris, 2006). However, Norris et al. (2003) describe the possibly game-changing insight that one must be careful to distinguish between *online feedback* (as in TRACE) and *feedback for learning* (as in backpropagation). They argue that feedback for learning provides the necessary basis for precompiling context-sensitivity into forward connection weights, and suggest that if it turns out that online feedback exists, it may only be an epiphenomenon of the need for feedback for learning. Mirman, McClelland and Holt (2006b) note that both sorts of feedback are a natural consequence of the assumptions of interactive architectures. All the same, this interesting distinction may be the key to resolving the debate (Magnuson, 2008a).

4.2. Lexically-mediated phoneme inhibition.

A recurring theme among criticisms of feedback is that it would cause distorted or inaccurate perception at pre-lexical levels. Massaro (1989) argued that lexical feedback distorts the representation at the phoneme layer, causing TRACE to fail to fit data from experiments that separately manipulate auditory and lexical support for the identity of a phoneme. However, subsequent work showed that if intrinsic variability is implemented, feedback does not distort pre-lexical processing (McClelland, 1991), and an extension of this work proved that interactive models can implement optimal Bayesian inference for combining uncertain information from independent sources (Movellan & McClelland, 2001).

A related prediction is that if the acoustic input contains a lexically inconsistent phoneme (for example, if the /k/ in “arsenic” is replaced with /t/ to make “arsenit”), lexical feedback would cause a delay in recognition of the acoustically present phoneme. Two sets of experiments failed to find evidence of lexically-induced delays in phoneme recognition (Frauenfelder et al., 1990; Wurm & Samuel, 1997), providing a key motivation for the development of the autonomous Merge model (Norris et al., 2000). Mirman et al. (2005) showed that these experiments had conflated the manipulation designed to show lexical inhibition effects

with the lexical status and neighborhood structure of target items at the point of the lexically inconsistent phoneme target. The TRACE model predicted lexical inhibition when these factors were controlled, but not under the previously tested conditions, and behavioral tests were consistent with these predictions. Thus, lexical feedback can slow phoneme recognition.

Proponents of the autonomous view have argued that models with lexical feedback would "hallucinate" lexically consistent phonemes that were not present in the input (Norris et al., 2000; Norris & McQueen, 2008). This overstates the potential for hallucination in TRACE (as the "trace" preserves details of malformed input, and model behavior differs greatly given well- and malformed input; McClelland & Elman, 1986, e.g., Figures 7-11). In addition, the "hallucination" claim is typically described as a thought-experiment that falsifies interactive feedback, but this underestimates actual human misperception: in lexical inhibition tests (Mirman et al., 2005), listeners exhibited a tendency toward lexically-induced misperception and this finding is consistent with other contextually-appropriate but illusory perceptions of speech such as failures to detect mispronunciations (Cole, 1973; Marslen-Wilson & Welsh, 1978), hearing noise-replaced phonemes ("phoneme restoration": Samuel, 1981; 1996; 1997; Warren, 1970), and similar findings from other modalities, such as illusory visual contours (Lee & Nguyen, 2001). In sum, the pattern of phoneme identification phenomena in the literature, including lexically-induced delays and errors, is consistent with direct feedback from lexical to pre-lexical processing.

4.3 Lessons from the feedback debate.

The feedback debate continues with researchers on both sides providing new behavioral and computational arguments supporting their view (McClelland et al., 2006; McQueen et al., 2006; Mirman et al., 2006a). Nonetheless, the debate illustrates the critical two-way connection between model simulations and behavioral data: simulations need to fit the behavioral data and make predictions for new behavioral experiments. For this connection to work, simulation materials and linking hypotheses need to be matched to behavioral experiment materials and task constraints and intuitive model predictions need to be tested with empirical simulations. In addition, resolving the debate may require integration with other domains of cognitive science (e.g., theoretical neuroscience: Magnuson, 2008a, Friston, 2003) and broader scope analyses (e.g., the importance of interactive feedback for learning).

5. Crucial questions and directions for progress

Current computational models of SWR theories require assumptions about the input and output and implementations of three core principles: multiple activation, similarity and priors, and competition. Progress may require us to reconsider where SWR begins and ends. SWR can be construed narrowly, as mapping strings of phonemes onto sound forms associated with words, or as broadly as mapping from the acoustic signal to a comprehensive set of phonological, grammatical, and semantic characteristics as part of the processes of recognizing larger structures like sentences (cf. Dahan & Magnuson, 2006). Whether you adopt a narrow view, broad view, or something in between has dramatic implications for your processing theory. The conventional view is that adopting the simplifying assumptions of the narrow view allows us to break off a tractable piece of the problem. But seemingly minor simplifying assumptions may actually complicate things, because they remove potentially constraining information.

Consider the *embedded word problem*. Most words are embedded in other words, and/or have words embedded within them (depending on dialect, *cat, at, a, cattle, law* and *log* are

embedded in a phonemic transcription of *catalog*), suggesting that models of SWR must somehow inhibit recognition of embedded words. The problem is much less extreme when one considers potential subphonemic cues such as durational differences between short and long words. For example, the syllable /haem/ is longer in the word *ham* than in *hamster*. Salverda et al. (2003) used eye tracking to measure lexical activation and competition and found that subjects were exquisitely sensitive to vowel duration differences of only about 15 msecs (see Davis et al., 2002, for converging results from priming studies), suggesting such subphonemic cues may mitigate (but not obviate) the embedding problem. Thus, while adopting the narrow view of SWR may allow traction on significant parts of the problem, it may simultaneously complicate the problem by ignoring useful information. The same holds in the opposite direction: limiting the scope of SWR to phonological form recognition ignores syntactic, semantic, and pragmatic knowledge that could potentially constrain word recognition. Similarly, eschewing production constraints, as well as learning and developmental trajectories leaves a more tractable problem, but at the peril of missing, for example, ways in which seeming puzzles of adult processing might emerge in unanticipated fashion from developmental pressures (MacDonald, 1999).

In our view, the greatest potential for progress in modeling SWR is in taking increasingly broader views: upstream (by working towards models that operate on raw speech), downstream (by connecting the output of current SWR models with higher order linguistic and cognitive structures), and developmentally. Current debates, like the feedback debate, have little consequence for broad-view models; the differences between models are modest and may disappear (or be amplified) as we grapple with greater realism. The model of Plaut and Kello (1999), with its realistic inputs, perception-production connections, and developmental approach, stands out as a promising example of how the field might proceed towards these goals.

Author notes

Preparation of this chapter was supported by NIDCD grant DC-005765 and NSF grant 0748684 to JSM, and NICHD grants F32HD052364 to DM and HD-01994 to Haskins Laboratories.

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.

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

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

Cole, R. A. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception & Psychophysics, 1*, 153-156.

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

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

Dahan, D., & Magnuson, J. S. (2006). Spoken-word recognition. In M. A. Gernsbacher & M. J. Traxler (Eds.), *Handbook of Psycholinguistics* (pp. 249-283). Elsevier.

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

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

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

Ellis, R., & Humphreys, G. W. (1999). *Connectionist psychology: A text with readings*. Hove,England: Psychology Press/Taylor & Francis (UK).

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179-211.

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.

Fowler, C. & Magnuson, J. S. (this volume). Speech perception. In M. Spivey, K. McRae, & M. Joanisse (Eds.), *The Cambridge Handbook of Psycholinguistics*. Cambridge University Press.

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

Frauenfelder, U. H., Segui, J., & Dijkstra, T. (1990). Lexical effects in phonemic processing: Facilitatory or inhibitory? *Journal of Experimental Psychology: Human Perception & Performance, 16*(1), 77-91.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks, 16,* 1325-1352.

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. (1999). Ambiguity, competition, and blending in spoken word recognition. *Cognitive Science, 23*, 439-462.

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

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

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

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

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate

speech categorization. *Journal of Experimental Psychology: Human Perception and Performance, 23*(2), 481-503.

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

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

Jacobs, Arthur M. and Grainger, Jonathan (1994) Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance, 20(6)*, 1311-1334.

Kello, C. T., & Plaut, D. C. (2003). Strategic control over rate of processing in word reading: A computational investigation of the tempo-naming task. *Journal of Memory and Language, 48*, 207-232.

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

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

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*(2), 262-268.

Lee, T. S., & Nguyen, M. (2001). Dynamics of subjective contour formation in early visual cortex. *Proceedings of the National Academy of Sciences, 98*(4), 1907-1977.

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

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

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

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

MacDonald, M. C. (1999). Distributional information in language comprehension, production, and acquisition; Three puzzles and a moral. In B. MacWhinney (Ed.), The emergence of language (pp. 177-196). Mahwah, NJ: Erlbaum.

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

Magnuson, J. S. (2008b). Generating individual eye movement behavior from central tendency models of spoken word recognition. Technical Report, University of Connecticut. http://magnuson.psy.uconn.edu/pub.html.

Magnuson, J. S., Dahan, D., & Tanenhaus, M. K. (2001). On the interpretation of computational models: The case of TRACE. In J. S. Magnuson and K.M. Crosswhite (Eds.), University of Rochester Working Papers in the Language Sciences, 2 (1), 71 - 91.

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

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003b). Lexical

effects on compensation for coarticulation: A tale of two systems? *Cognitive Science, 27*, 801-805.

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

Magnuson, J. S., Tanenhaus, M. K., and Aslin, R. N. (2000). Simple recurrent networks and competition effects in spoken word recognition. *University of Rochester Working Papers in the Language Science, 1,* 56-71.

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

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

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

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

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

Marslen-Wilson, W., & Warren, P. (1994). Levels of Perceptual Representation and Process in Lexical Access: Words Phonemes and Features. *Psychological Review, 101*, 653-675.

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

Marslen-Wilson, W.D. (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.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology, 21*(3), 398-421.

Masson, M. E. J. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(1), 3-23.

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

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

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

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

McLeod, P., Plunkett, K. & Rolls, E. T. (1998). *Introduction to Connectionist Modeling of Cognitive Processes*. Oxford: Oxford University Press.

McQueen, J. M. (2003). The ghost of Christmas future: didn't Scrooge learn to be good? Commentary on Magnuson, McMurray, Tanenhaus, and Aslin (2003). *Cognitive Science, 27*(5), 795-799.

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

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*(3), 424-443.

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

Mirman, D., McClelland, J.L., and Holt, L.L. (2006b). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review, 13(6),* 958-965.

Morton, J. (1969) The integration of information in word recognition. *Psychological Review, 76,* 165-178.

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

Norris, D. (1990). A dynamic-net model of human speech recognition. In G.T.M. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*, pp. 87-104. Cambridge: MIT.

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

Norris, D. (2005) How do computational models help us build better theories? In A. Cutler, (Ed.) *Twenty-First Century Psycholinguistics: Four Cornerstones*. Mahwah, NJ: Lawrence Erlbaum.

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

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

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

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

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

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA,US: The MIT Press.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review, 113(1),* 57-83.

Pitt, M.A., Myung, J.I., & Altieri, N. (2007). Modeling the word recognition data of Vitevitch and Luce (1998): Is it ARTful? *Psychonomic Bulletin & Review, 14,* 442-448.

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

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*(1), 56-115.

Plunkett, K. & Elman, J. L. (1997) Exercises in Rethinking Innateness: A Handbook for Connectionist Simulations. Cambridge, MA: MIT Press.

Protopapas, A. (1999). Connectionist modeling of speech perception. *Psychological*

Bulletin, 125(4), 410-436.

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press.

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

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

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

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

Samuel, A. G. (this volume). Spoken word recognition. In M. Spivey, K. McRae, & M. Joanisse (Eds.), *The Cambridge Handbook of Psycholinguistics*. Cambridge University Press.

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

Scharenborg, O., Norris, D., ten Bosch, L. & McQueen, J. (2005) How should a speech recognizer work? *Cognitive Science, 29*, 867-918.

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

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

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

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

Trueswell, J. C., & Tanenhaus, M. K. (Eds.). (1994). *Toward a lexicalist framework of constraint-based syntactic ambiguity resolution*. Hillsdale, NJ,England: Lawrence Erlbaum Associates, Inc.

Van Berkum, J.J.A., Brown, C.M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. Journal of Experimental Psychology: Learning, Memory, & Cognition, 31(3), 443-467.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*, 392-393.

Wurm, L. H., & Samuel, A. G. (1997). Lexical inhibition and attentional allocation during speech perception: Evidence from phoneme monitoring. *Journal of Memory & Language, 36*(2), 165-187.

Appendix: Modeling tools.

This table lists tools useful for modeling spoken word recognition. The list is ordered by ease of use. Many more tools exist (such as the neural network toolbox for Matlab).

Tool	Description and URL
tlearn	Simple yet powerful simulator for feedforward and (simple) recurrent networks. No programming experience required. Batch processing possible with X11 version or scripting tools. Useful in conjunction with Plunkett & Elman (1997) and/or McLeod, Plunkett, & Rolls (1998). http://crl.ucsd.edu/innate
lens	Doug Rohde's "light, efficient neural simulator." Flexible tool for very wide range of neural networks. Graphical user interface. Tcl/tk interface makes basic programming skills useful, but not necessary. http://tedlab.mit.edu/~dr/Lens
Emergent	Very powerful tool for "parallel distributed processing" modeling, ranging from high-level cognitive models to neuronal models. Steep learning curve, but incredibly flexible. See O'Reilly and Munakata (2000). http://grey.colorado.edu/emergent/index.php/Main_Page
TRACE	jTRACE: Platform-independent reimplementation of the TRACE model in Java. Includes graphical user interface, analysis, graphing, scripting, and sharing tools. No programming experience required. See Strauss et al. (2007): http://magnuson.psy.uconn.edu/jtrace HebbTRACE: Original TRACE code (written in C), revised and augmented with Hebbian learning (Mirman et al., 2006b): http://magnuson.psy.uconn.edu/mirman/research/HebbTRACE.zip Mark Pitt provides a version of the original code that he has modified slightly and augmented with tools that facilitate simulation and analysis. http://lpl.psy.ohio-state.edu/software.html
ART	Mark Pitt provides Matlab code and descriptions of the version of ARTPHONE used by Pitt, Myung, and Altieri (2007). http://lpl.psy.ohio-state.edu/software.html

Endnotes

[i] We will stretch "computational model" to mean "formal model:" any formalism that describes a mapping. This definition is broad enough to include non-implemented descriptions of such a mapping process (verbal-algorithmic), as well as simple mathematical models. Note that there is also unfortunate potential for confusion over the common usage of "computational model" to refer to this range of approaches, and the most abstract of Marr's (1982) levels of information processing theories. A theory at his "computational level" describes a computed function in terms of input, output, and constraints on the mapping between them, in contrast to theories at the "algorithmic" and "implementational" levels (akin roughly to software and hardware, respectively). Mathematical models commonly reside at Marr's computational level, while verbal-algorithmic and simulation models commonly reside at his algorithmic level.

[ii] While PARSYN is *consistent* with NAM, as we will discuss, NAM's power is in its simplicity and remove from processing details as a choice model. We see PARSYN as complementary to NAM rather than a direct extension. While we are using the label "mathematical" to distinguish NAM from verbal-algorithmic and simulation models, note that this is a weak distinction, as complete mathematical descriptions of processing models may or may not be tractable. The key point here is the simplicity of the model, and its relation to Marr's (1982) computational level of information processing theories.

[iii] Technically, we are discussing "TRACE II;" TRACE I (Elman & McClelland, 1986) was focused on the speech-to-phoneme side of the model, but was never linked to TRACE II.

[iv] Shortlist is often incorrectly described as having a single shortlist, with all items inhibiting each other. Instead, there are shortlists aligned at each input position (making the potential size of the interactive activation network *sl:* s = maximum size of each shortlist, which is 30 by default; l = phonemic length of the input). Only items that overlap in time inhibit each other. See Figure 3.

[v] Similarly, Scharenborg, Norris, ten Bosch, and McQueen (2005) have proposed using automatic speech recognition (ASR) mechanisms for lexical search. As with SRNs, since ASR mechanisms themselves predict lexical competition (e.g., via rank ordered hypotheses), adding a competition network (Shortlist) to an ASR front-end may just account for processing delays, rather than providing a useful function unavailable in the ASR framework.

[vi] This is no small burden when a model has been shown to account for a wide range of results. However, tools like jTRACE (Strauss, Harris, & Magnuson, 2007) and others listed in the appendix allow one to automate large numbers of simulations in order to explore the robustness of previous simulations throughout parameter spaces.

[vii] The best fits used a parameter that changed over time, to reflect greater confidence as bottom-up evidence increased. With this parameter (k) set to a constant value of 7, competitor fits were reduced slightly. In later work, a constant value of 7 provided excellent fits (Dahan, Magnuson, & Tanenhaus, 2001; Dahan, Magnuson, Tanenhaus, & Hogan, 2001).

[viii] Dahan et al. (2001a, 2001b) extended these linking hypotheses to studies of frequency and subcategorical mismatch. The simple assumptions about the role of the visual display allow accurate predictions of changes in target fixations depending on whether a competitor is present in the display. Norris (2005) suggests that computing response probabilities corresponds to predicting that subjects' eyes instantaneously flit between objects (that is, that

each trial must have the same continuous form as the central tendency). However, in choice theory, a response *probability* implies a distribution of responses. Magnuson (2008b) provides simulations demonstrating that a 1-parameter stochastic eye movement model quickly recovers the underlying distribution.

[ix] PSP tools are available from:
http://faculty.psy.ohio-state.edu/myung/personal/PSP_PAGE.html.