

3 **Considering the junction model of lexical processing**

Christopher T. Kello

Department of Psychology, George Mason University, USA

The work presented in this chapter arose from two questions that may at first appear unrelated: should we accept the premise that skilled word reading is supported by lexical and sublexical routes of processing, and how can we model the learning of orthographic and phonological representations for multisyllabic words? The field as a whole has answered “yes” to the first question, but is mostly silent with regard to the second. Here I will consider an alternative to lexical and sublexical routes, one in which these are modes of processing instead of components of processing. My colleagues and I have pursued this idea by exploring theories of lexical processing in which orthographic, phonological, and semantic codes are all mediated by one level of representation. In our pursuit, we were forced to confront the long-standing problem of modelling multisyllabic words. This chapter is an account of our efforts that have recently culminated in the junction model of lexical processing.

The story begins with yet another question: how does speech fit into theories of word reading? It is only common sense that written language processes are built upon spoken language processes. Spoken language takes precedence in all senses of the word. It is during speech acquisition that the phonological, morphological, and semantic structures of words are learned. These structures are learned as a bridge between speech signals (acoustic, optical, and articulatory) and the rest of the language system. When orthographic structures are introduced, they must be learned in a way that fits with the bridge already in place.

So how are orthographic inputs mapped onto spoken language processes in current theories of word reading? An answer can be found in diagrams of the two major current frameworks, shown in Figure 3.1. In the dual-route cascaded (DRC) model (Coltheart, Curtis, Atkins, & Haller, 1993; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001), orthographic inputs connect with the spoken language system via two routes of processing, one lexical and the other sublexical. The sublexical processes break down the orthographic forms of words into sequences of graphemes, and a rule is used to essentially map each grapheme onto a corresponding phoneme. Phonemes would presumably be learned during spoken language acquisition, and should therefore be

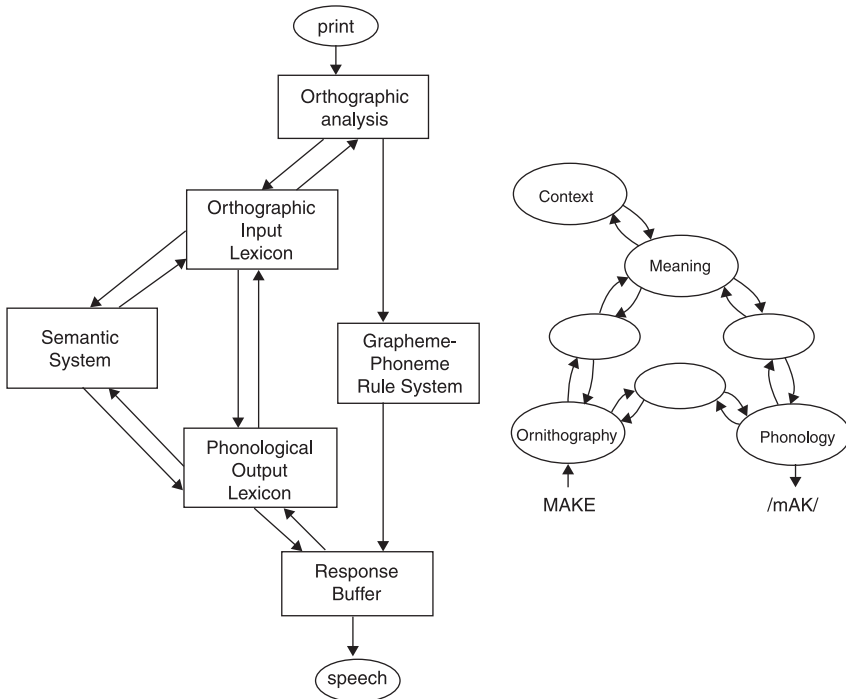


Figure 3.1 Diagrams of two theories of the lexical processing system. The one on the left is taken from Coltheart et al. (2001), and the one on the right is taken from Seidenberg and McClelland (1989); both are reprinted with permission. Lexical and sublexical routes of processing can be found in both theories.

considered as part of the spoken language system. The lexical processes associate each orthographic word form as a whole with semantic and phonological representations that would also be learned during spoken language acquisition.

The answer is similar for the triangle framework of lexical processing (Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989). Orthographic inputs are mapped onto phonological and semantic representations via two corresponding sets of hidden units. The orthography-phonology mapping is sublexical in nature because in most languages there tend to be systematic relations between orthographic and phonological units that are smaller than the word. The orthography-semantic mapping is lexical in nature because the relation between orthographic units and meaning is mostly unsystematic within individual morphemes (thus, in English, the letter “d” bears no relation to the meaning of DOG).

Why is it hypothesized in these frameworks that orthographic inputs connect with the spoken language system via two separate routes, one lexical and

the other sublexical? One reason is historical: researchers at one time believed that the functions performed by skilled readers required these two routes. On the one hand, skilled readers can pronounce words like PINT and HAVE even though they do not conform to the systematic relations between orthography and phonology. This ability suggests that a lexical route is necessary to store the pronunciations of such irregular words individually. On the other hand, skilled readers can give plausible pronunciations for nonwords like PILT and HOVE. This ability suggests that a sublexical route is necessary to assemble novel pronunciations from grapheme-phoneme correspondences or the like.

As it turns out, the ability to read aloud both irregular words and nonwords does not necessarily require a lexical and sublexical route. Glushko (1979) explained how lexical representations might be used to generate nonword pronunciations via a process of analogy. For instance the pronunciation of PILT might be computed as a blend of words like SILT and PILL. Such an analogy model would be able to generate pronunciations for both irregular words and nonwords via a single route of processing. Seidenberg and McClelland (1989) then showed how a single level of distributed representation in a connectionist model can be used to generate pronunciations for both kinds of items.

These demonstrations are valuable in delineating the space of viable theories, but most researchers still believe that orthographic inputs are connected to the spoken language system via two routes. The main reason is evidence for the dissociation of lexical and sublexical processes. This evidence has most notably come from surface versus phonological dyslexia in their acquired forms. Surface dyslexia is characterized by poor reading of irregular words with relatively intact reading of nonwords (Behrmann & Bub, 1992). Phonological dyslexia is characterized as the opposite, thereby forming a double dissociation (Funnell, 1983). Double dissociations are usually interpreted as evidence for separable processes, and this dissociation is no exception. How could surface and phonological dyslexia occur without a lexical route and a sublexical route? The question seems hard to answer, not to mention other arguments for lexical and sublexical routes that have been made on the basis of evidence for strategic control of the two routes (Monsell, Patterson, Graham, Hughes, & Milroy, 1992; Zevin & Balota, 2000).

Considering a shared-route architecture

The idea of lexical and sublexical routes appears to have good empirical support. But let us again consider the interface of orthography with the spoken language system. Does it make good sense for orthography to interface along these two routes? The core competency of spoken language is to map between sound and meaning. We can probably all agree that processes and representations are learned during spoken language acquisition to support this mapping. Reading requires a mapping to sound and meaning. What

if orthography interfaced with the representations that mediate sound and meaning (Figure 3.2), rather than with sound and meaning themselves?

This idea has a certain efficiency to it. To start with, only one route of processing would be needed between orthography and the spoken language system. The loss of a route seems more efficient in principle, but there is a more specific advantage in this case. Orthographic inputs would be interfaced with what may be construed as morphological representations, because morphology is at the intersection of phonology and semantics (Plaut & Gonnerman, 2000). If so, morphological structure would not need to be duplicated across two routes, as it is with lexical and sublexical routes. Having two routes leads to the duplication of morphology because the relation between semantics and phonology is structured similarly to that between semantics and orthography.

So far so good, but there is an obvious question. What about the evidence

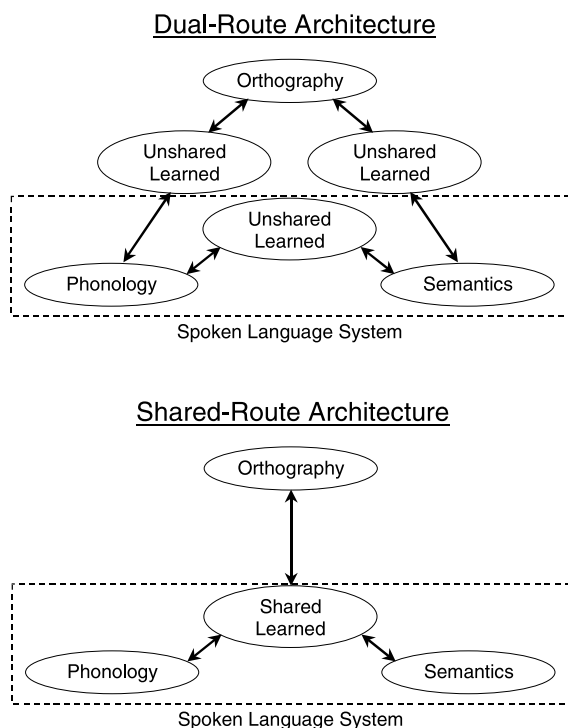


Figure 3.2 Diagrams comparing dual-route and shared-route architectures. The well-known triangle model is used to illustrate the theoretical commitments that apply more generally to dual-route architectures. The shared-route architecture is implemented in this chapter as the junction model of lexical processing. The outlined portions represent the spoken language parts of these architectures, which are presumably in place prior to the development of the written language parts.

for a dissociation between lexical and sublexical processes? Does this evidence already rule out a shared-route architecture? Not necessarily. For one thing, Patterson and her colleagues have proposed that surface and phonological dyslexia may be explained by damage to semantic and phonological representations, respectively (Patterson & Hodges, 1992; Patterson & Marcel, 1992). The logic of this idea can be seen in the fact that, in the dual-route architecture, one terminal of the lexical route is semantics, and one terminal of the sublexical route is phonology. The idea has been criticized on the grounds that surface and phonological dyslexias do not always coincide with semantic and phonological impairments (Coltheart, 1996), but there have been counterarguments (Patterson & Lambon-Ralph, 1999). It is fair to say that the debate is undecided, which means that the shared-route verdict remains undecided.

The shared-route architecture gains even more credibility when we consider a second way to account for the lexical/sublexical dissociation. Kello and his colleagues (Kello, 2003; Kello, Sibley, & Plaut, 2005a; Sibley & Kello, 2004) recently showed how lexical and sublexical processing can be two modes of a single processing route. A control parameter termed “input gain” was manipulated in a series of demonstration models with only one route of processing. The single route was mediated by either localist or distributed representations. Input gain had the effect of scaling the net inputs to processing units. For both kinds of representation, low and high levels of input gain shifted the models between *regularity-based* and *item-based* modes of processing, respectively.

These modes resulted in a clear double dissociation, as shown in Figure 3.3. The models were given a mapping task for a corpus of items that mostly but not entirely conformed to a simple rule (the identity mapping). Mapping accuracy is plotted as a function of input gain for regular trained items (those conforming to the identity mapping), irregular trained items (those not conforming to the identity mapping), and untrained items. The pattern of results was the same for both localist and distributed models. At low input gain, performance was near perfect for regular and novel items, but nearly all of the irregular items were “regularized”, that is, incorrectly given the identity mapping. This performance profile is analogous to the defining characteristics of surface dyslexia. At high input gain, performance was highly accurate for all trained items, but poor for novel items. This profile is analogous to the defining characteristics of phonological dyslexia.

Input gain caused a double dissociation for both localist and distributed representations because it affected the scope of knowledge that was brought to bear on the mapping task. At low input gain, each input pattern was mapped according to a wide range of trained items with similar patterns. Irregularities were essentially averaged out of the mapping. At high input gain, each input pattern was mapped according to a much smaller range of similar items, hindering the generalization that relies on regularities spanning across multiple items.

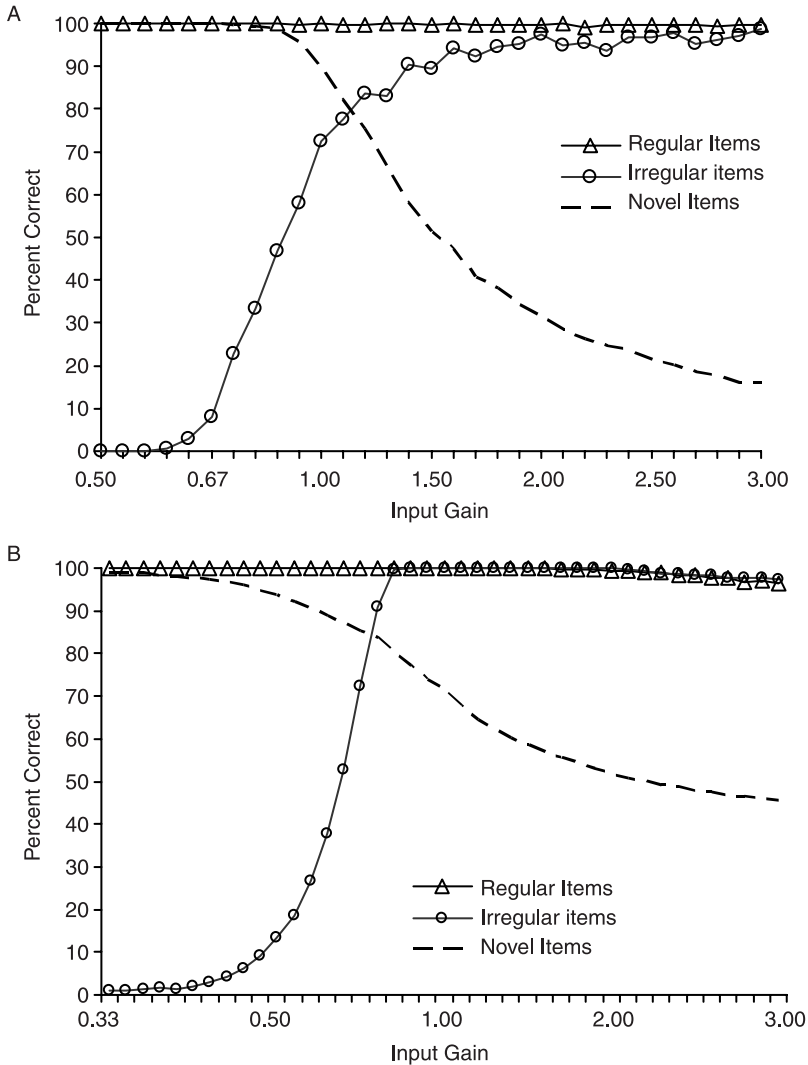


Figure 3.3 Performance of localist (A) and distributed (B) models as a function of item type and input gain (reprinted with permission from Kello et al., 2005a).

Input gain demonstrates a mechanism by which lexical (item-based) and sublexical (regularity-based) processing may be two modes of operation in a shared-route architecture. Surface and phonological dyslexia could therefore result from trauma or developmental abnormalities that restrict the flexibility of processing to a single mode. While there are well-established neural correlates to input gain (Fellous & Linster, 1998), there are currently no data to determine whether such correlates may play a role in surface or phonological dyslexia.

A hint of evidence in tempo-naming

At this point in the argument, we can say that there are principled reasons for considering a shared route of processing between orthography, phonology, and semantics. We can also say that a shared-route architecture is at least viable in the face of dissociations between lexical and sublexical processes. But is there any evidence in favour of such a route? In fact, the idea was pursued in the first place because of observed naming errors that were readily accommodated by a shared-route architecture, but not by dual-route architectures.

The naming errors came from the tempo-naming task that was introduced by Kello and Plaut (Kello, 2004; Kello & Plaut, 2000). In this task, an audiovisual metronome is used to control the amount of time between the onset of a letter string and the initiation of a naming response. The metronome is played for a set number of beats on each trial, and a letter string is displayed on the final beat. Participants are instructed to name the letter string such that their response is timed with what would be the next beat of the metronome. Feedback is given at the end of each trial to indicate whether the response was initiated early or late relative to the metronome. Fast tempos can be used to induce a speed/accuracy trade-off because they reduce the amount of time between stimulus onset and the cue to respond.

Numerous experiments have shown that fast tempos lead to four basic categories of errors: articulatory errors (e.g. slurs and mis-starts), regularizations (e.g. PINT pronounced to rhyme with MINT), lexicalizations (e.g. BOARD pronounced as a similar word like BROAD), and nonword errors (e.g. STINT pronounced as a similar nonword like STIT). The question is, which kinds of errors should become more prevalent under time pressure, given a separation of lexical and sublexical routes? To answer this question, Kello and Plaut (2000) ran simulations of the tempo-naming task with the triangle and DRC models of lexical processing. They found that both models generated a comparable increase in both regularization and lexicalization errors under time pressure. The time pressure caused by faster tempos was simulated by sampling the models' outputs at earlier points in the time course of processing relative to stimulus onset. In another simulation, Kello and Plaut (2003) found the same result for the triangle model when time pressure was simulated by increasing the rate of processing.

The tempo-naming experiments have yielded a different result. As tempos increase, the rate of regularization errors has been shown to remain constant while the rates of other types of errors, including lexicalization errors, increase. The lack of an increase in regularization errors means that time pressure in the tempo-naming task does not lead to a misapplication of sublexical spelling-sound correspondences. But this is precisely what happened when time pressure was implemented in the triangle and DRC models. So the question now becomes, how can time pressure be implemented in a

model of lexical processing such that it does not lead to a misapplication of spelling-sound correspondences?

One possible answer can be found by delving into the reason behind the effect of time pressure in the triangle model. As mentioned earlier, the mapping between orthography and phonology is much more systematic (due to spelling-sound correspondences) than the mapping between semantics and either orthography or phonology, at least in English. Distributed representations are more adept at processing systematic mappings than arbitrary mappings. This is because, by default, similar inputs generate similar distributed representations, and similar distributed representations generate similar outputs. As a result of this adeptness, the relatively direct mapping between orthography and phonology is computed more quickly in the triangle model, compared with the same mapping that is mediated by semantics (Van Orden, Bosman, Goldinger, & Farrar, 1997). The consequence is that spelling-sound correspondences represented in the direct mapping lead to increases in regularization errors when time pressure is simulated.

One way to avoid the increases in regularization errors is to simply remove the direct mapping from orthography to phonology. The shared-route architecture does this by essentially combining the route from orthography to phonology with the route from orthography to semantics. Kello and Plaut (2003) reported a simulation of the shared-route architecture to show that it could in fact provide a closer account of the observed errors in tempo-naming than the triangle model. They also showed that the shared-route architecture can account for the hallmark effects of printed frequency and spelling-sound regularity in word reading. It should be mentioned that the dual-route architecture may ultimately be reconciled with the tempo-naming data, but the point here is that it was the challenge of the tempo-naming data that led us to reconsider the dual-route architecture.

Large-scale modelling and the problem of multisyllabic words

I have now laid out the logical, empirical, and computational arguments in favour of a shared-route architecture. These arguments may not sway a proponent of dual-route architectures, but they should at least provide justification for further pursuit. The model presented by Kello and Plaut (2003) was a good start, but it was small (a corpus of 470 words) in comparison to the DRC and triangle models that have been reported. The precedent set by these models calls for a large-scale simulation of the shared-route architecture. The term “large-scale” in the domain of lexical processing currently refers to models that contain roughly between 3000 and 8000 words.

But if we fully embrace the notion that large-scale simulations provide strong tests of theories of lexical processing, we must go beyond the extant models because they are restricted to processing monosyllabic words. One might at first think that the inclusion of multisyllabic words is only a matter of scaling up the current models, but in fact we are forced to confront a

fundamental issue that was heretofore marginalized: how does one represent variable-length sequences of letters or sounds? Moreover, how are such representations learned? Current models provide inadequate answers to the first question, and no answers to the second question.

The representation of multisyllabic word forms is difficult because it engenders a binding problem (Rosenblatt, 1961; von der Malsburg, 1981) in which sounds or letters must be bound to their positions in the word form. Information must be learned and represented about elements with respect to their positions, as well as independently of their positions. Slot-based codes have been used in most models, but these engender the alignment and dispersion problems that were explained by Plaut et al. (1996). The “Wickelfeatures” that were used by Seidenberg and McClelland engender a different set of problems (see Pinker & Prince, 1988). Perhaps a scheme of representation can be engineered to overcome the known problems of orthographic and phonological representation (for instance, see *C. J. Davis*, this volume), but ultimately one must come to grips with the fact that these representations are learned, not engineered.

This fact aside, it is important to recognize the progress that has been made with the extant models using engineered representations of monosyllabic words. There were and still are many theoretical questions for which the problem of multisyllabic words can be put off until a later date. It also must be noted that there have been very few data on multisyllabic words to test the models against. However, the recent development of the Elexicon database has drastically changed the size of the playing field (Balota et al., 2002). The database currently contains lexical-decision and word-naming data for over 30,000 English words, and most of these words are multisyllabic. These data constitute a vast source of evidence for testing models of lexical processing, but they cannot fully be put to use until the problem of multisyllabic words is addressed.

The upcoming junction model is a first-pass effort at testing the shared-route architecture against the Elexicon database. In this model, the learning of multisyllabic words was accomplished by drawing upon earlier modelling work by Plaut and Kello (1999). They introduced a theory of phonological development in which the phonological representations of words are learned in the service of speech comprehension, production, and imitation. The theory was implemented in a connectionist model in which phonological representations were learned through three language tasks: (1) the integration of variable-length acoustic sequences into fixed-width representations, and the use of those fixed-width representations to (2) activate semantic knowledge (i.e. speech comprehension) and (3) generate variable-length sequences of articulatory outputs (i.e. speech production).

The means to learn of multisyllabic word forms lies in this model’s ability to learn fixed-width representations of variable-length sequences. Mono- and multisyllabic word forms vary widely in their numbers of letters and sounds. By converting such variable-length sequences into fixed-width representations,

a model can leverage the benefits of distributed representation and processing. Such representations would provide a common basis for relating word forms of different lengths, and for generalizing to novel word forms of varying lengths. Information about letters and sounds would be learned both within and across their positions in orthographic and phonological word forms.

Kello and his colleagues (Kello, Sibley, & Colombi, 2004; Kello, Sibley, Plaut, & Elman, submitted, 2005b) recently provided a proof-of-concept that the methods developed by Plaut and Kello (1999) can indeed be used to learn representations of both mono- and multisyllabic word forms. The basic innovation was to extend the simple recurrent network architecture (SRN) (Elman, 1990; Jordan, 1986) to learn fixed-width representations in the dual service of both encoding and decoding variable-length sequences. SRNs have been used most often in models of language processing to learn the transitional probabilities of sequences by training them to predict each subsequent element from previous elements. While transitional probabilities can be used naturally as assays of grammatical learning (Christiansen, Allen, & Seidenberg, 1998; Cleeremans, Servan-Schreiber, & McClelland, 1989), they do not translate directly into a method for learning fixed-width representations. The problem is that the task of learning transitional probabilities requires the network to hold *only* enough information about a sequence to predict its next element. Thus, there is no pressure to encode or decode an entire sequence.

The *sequence encoder* developed by Kello and his colleagues creates the necessary pressure by coupling two SRNs in order to simulate what is essentially the imitation of variable-length sequences (Figure 3.4). An encoder SRN is trained to generate a fixed-width representation of a given sequence, and a decoder SRN is trained to regenerate that sequence from the fixed-width representation. At the beginning of training, representations generated by the encoder SRN are mostly arbitrary, and the decoder SRN is pressured

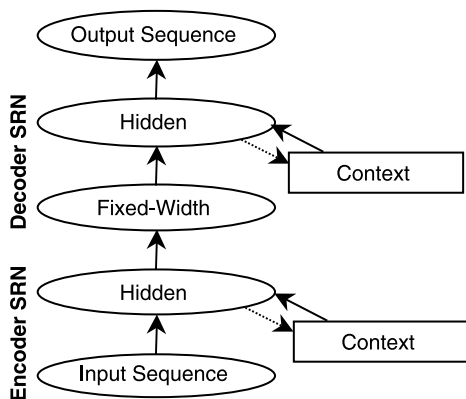


Figure 3.4 Diagram of the sequence encoder architecture.

to reproduce the input sequences from these arbitrary representations. As learning progresses, weight changes in the encoder SRN adapt the fixed-width representations to error signals that come from the decoder SRN. At the same time, weight changes in the decoder SRN adapt to the fixed-width representations that are generated by the encoder SRN.

Simulations with small, artificial sets of sequences have shown that the sequence encoder is capable of learning to encode and decode variable-length sequences (Kello et al., 2004; submitted, 2005b). Learning also generalized very well to novel sequences. Analyses showed that the elements of a sequence were represented both with and without respect to their positions, where position is defined either in terms of rank order or relative to other elements in the sequence. Moreover, large-scale models have been built to learn orthographic and phonological representations for nearly 75,000 English words. These models were shown to generalize very well to legal pseudowords: 82% of over 71,000 phonological pseudoword forms were processed correctly, and 76% of over 74,000 orthographic pseudoword forms were processed correctly. Generalization to illegal nonwords was only 18% and 23%, respectively, which indicates that learning was shaped by the structures of English word forms. This shaping was also seen in that the errors made by the sequence encoder tended to create legal word forms. Space limitations prohibit a more detailed discussion of the sequence encoder, but the brief overview given here is meant only to motivate the sequence encoder's use in the upcoming model of lexical processing.

The last kind of representation to consider is semantics. At first, it may not be clear whether the sequence encoder is appropriate for simulating the learning of semantic representations, but it would be if one views such representations as being formed at the junction of perception and action. For instance, consider how the semantic aspects of a word like *BAT* are grounded in our experiences with bats and batting. It is plausible to hypothesize that, at some level of abstraction, our perceptual representations of bats and batting have structure that is systematically related to our action representations of bats and batting, e.g., the shape and weight of a bat captures something about how bats are perceived, as well as how they are acted upon. The principle of perception/action junctions would lead one to hypothesize representations that capture the common ground between perception and action that exists for any given concept denoted by a word. This idea has much in common with Gibson's affordances (Gibson, 1979). If one goes on to say that even abstract concepts are grounded in experience, the idea also has much in common with Lakoff and Johnson's conceptual metaphors (Lakoff & Johnson, 1980), and the recent movement toward embodied cognition (Wilson, 2002).

As appealing as the idea may sound, there are many questions that remain to be answered. What would the sequences of perceptual inputs and action outputs consist of? How would one handle polysemy and the contextual flexibility of meaning? Could this embodied idea cover even the most abstract

aspects of semantics? And so on. In the upcoming model, these questions are set aside by using semantic representations built from textual co-occurrence statistics as proxies for junction representations. While one could raise questions about the validity of using textual co-occurrences to derive semantic representations (see Harm & Seidenberg, 2004), they are at least sufficient approximations given the current knowledge of lexical semantics.

Putting it all together: A large-scale pilot of the junction model

We now have the motivation and means of implementing a large-scale model of lexical processing by a shared-route architecture. The model presented here was built from a corpus of 45,273 English words, and tested against response times for over 30,000 of those words. The simulated response times were compared against response times in the Elexicon database (Balota et al., 2002). The proportions of explained variance are benchmarked against regression models with word frequency, length, and articulatory factors as the predictors. Results from the pilot model are also benchmarked against results from Plaut et al.'s (1996) simulations and the DRC model of word recognition (Coltheart et al., 2001). The chapter closes with a discussion of how the pilot model can be improved and extended to more fully address the range of phenomena relevant to lexical processing.

Model architecture

The junction model implemented here is diagrammed in Figure 3.5. The orthographic, phonological, and semantic representations were first constructed separate from each other, and then bound by a set of mediating

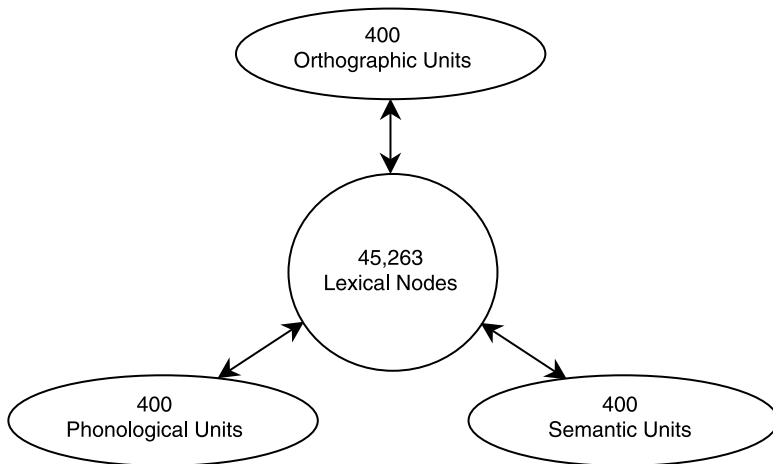


Figure 3.5 Architecture of the junction model. Arrows denote full connectivity.

representations. The pilot model was constructed piecemeal in order to simplify matters. The theory actually calls for spoken language acquisition to be simulated as the learning and binding of phonological and semantic representations. Then the learning of orthographic representations would be gradually integrated with the spoken language system. For the sake of simplicity, this developmental side to the model was not implemented here, but is planned for future simulations.

Orthographic, phonological, and semantic representations

Sequence encoders were used to learn an orthographic and a phonological representation for each of the 45,263 words in the training corpus. The training corpus was comprised of the intersection of the Carnegie Mellon University (CMU) pronunciation dictionary, the COALS database of word co-occurrences (Rohde, Gonnerman, & Plaut, in preparation, 2006), and the Microsoft spelling dictionary. The orthographic and phonological representations were learned with two pairs of sequence encoders, one pair for orthography and one pair for phonology. The orthography pair learned to encode and decode sequences of capital letters that comprised English words. The phonology pair learned to encode and decode sequences of phonemes. Each letter was coded by a pattern of 10 binary features, such as “has a left vertical line” and “has a curve”. Each phoneme was coded by 10 phonetic features, such as frication and voicing, plus two additional features to code for lexical stress on vowels.

The pairs were used to create two hierarchical models of two stages each. The first stage was used to learn to encode and decode sequences of letters or phonemes into vowel groups that roughly corresponded to syllables. The second stage was used to learn to encode and decode sequences of vowel groups into words, using 400 hidden units for the fixed-width representations. This two-stage method was used to reduce the number of elements that any one sequence encoder was required to process. The two-stage method was a simplification that helped make progress toward the larger effort of building a large-scale model of lexical processing. As mentioned earlier, more recent work with sequence encoders has shown that orthographic and phonological representations of words can be learned in a single stage (Kello et al., submitted, 2005b).

At the end of training, the sequence encoders could correctly encode and decode over 99% of the trained words. Two 400-bit codes were learned for each word, one orthographic code and one phonological code. The similarity structure of these codes is crudely illustrated in Figure 3.6. Six example words are listed along with their closest neighbours, three for orthography and three for phonology. Neighbours are listed in order of similarity as measured by the number of bits in common with each example word. The neighbours illustrate how similarity of the orthographic and phonological codes was driven by the overlap in letters or phonemes, semi-independently of their position.

		Orthography			Phonology		
		Deterministic	Jumping	Lice	Board	Hat	Institutionalization
Most Similar	determination	jumped	lace	bored	sat	institutionalizes	
	determinate	jumpers	vice	born	haut	institutionalizing	
	determining,	jumper	malice	borne	hath	institutionalize	
	determinative	bumping	device	bourne	seat	internationalization	
	determinations,	jumpy	nice	brought	set	institutionalized	
	determinable	ramping	sluice	bard	height	institutionally	
	determinedly	slumping	police,	barred	head	constitutionality	
	determinism,	lumping	face	bared	heat	institutional	
	determines	tamping	dice	gored	half	counterrevolutionary	
	determined	thumping	rice,	gourd	shad	internationalism	
Less Similar	determinants	pumping,	mice	mort	sec	unconstitutionally	
	determine	camping	slice	barn	heath	internationalized	
	paternalistic	stumping,	licorice	guard	haute	antidiscrimination	
	determinist	scrimping,	loci	marred	site	extraterritoriality	
	determinant	jumbo	solace,	morgue	sight	industrialization	
	mechanistic	dumping	pace,	bread	shit	denationalization	
	hedonistic	stomping	mace	bred	sheet	constitutional	
	pessimistic,	lamping	vide	chord	had	constitutionally	
	opportunistic			cord	cite	interdisciplinary recapitalization	

Figure 3.6 Illustration of the similarity structure of the orthographic and phonological representations learned by the sequencer models. Three example words for orthography are shown along with three example words for phonology, and below each example is a word list ordered from most similar to less similar.

Finally, one semantic representation was generated for each of the 45,273 words. Each representation was a 400-bit pattern derived from the COALS method of compiling co-occurrence statistics for words in texts (Rohde et al., in preparation, 2006). The method is similar to latent semantic analysis (Landauer & Dumais, 1997), and the statistics were culled from numerous and various text sources on the Internet. As noted earlier, the statistical extraction of co-occurrence statistics was a proxy for what would ideally be implemented as semantic junctions.

Mediating lexical nodes

This part of the pilot model may lead to some confusion and controversy without careful explanation. The defining aspect of the shared-route architecture is that a single level of representation is used to mediate the mapping between orthography, phonology, and semantics. In precursors to the junction model (Kello, 2003; Kello & Plaut, 2003), this mediating level of representation was learned by the back-propagating error that was incurred in mapping among the three types of codes. Back-propagation is a versatile learning algorithm, but it has certain biases and limitations. One of its biases is that systematic mappings are learned more easily (fewer training epochs

and/or hidden units) than unsystematic mappings. This bias works in favour of learning the mapping between orthography and phonology, but not between semantics and either orthography or phonology. All of these mappings must be supported by the mediating level of representation. In the precursor models, the bias was overcome by using a sufficient number of hidden units, and training for a sufficient number of epochs. But the corpus of words that was learned in those models was relatively small at 470 words. Using back-propagation to learn the mediating representations for over 45,000 words would take a prohibitive amount of time and computing power.

The alternative used in the current pilot model was to create lexical nodes and assign each one to represent one word in the corpus. Lexical nodes have no inherent bias in computing systematic versus unsystematic mappings. Because they can be prespecified, no training is required in order to use them. Theoretically, the use of lexical nodes in the pilot model makes it clear that the mediating level of representation in the shared-route architecture is a lexicon of sorts (see *Andrews*, this volume; *Taft*, this volume). This point is true regardless of whether the mediating level comprises lexical nodes or more distributed representations. But in using lexical nodes do we forfeit the benefits of learning and generalization that are conferred by distributed representations? The short answer is no.

The longer answer starts with the fact there are established algorithms for learning localist codes, such as the one developed by Grossberg in adaptive resonance theory (Grossberg, 1980). Such an algorithm will be used in future implementations, after other issues are worked out in earlier implementations. As for generalization, Kello et al. (2005a) showed how nonwords could, in principle, be processed by lexical nodes in the junction model. The simple idea is that the orthographic representation for a given nonword will be similar to the orthographic representations for similarly spelled words. This similarity gradient can be reflected in the activations of localist units (depending on their activation function), thereby creating a distribution of activation over the localist nodes for any given nonword. This distribution may in turn activate an appropriate phonological representation for the nonword, provided that the mapping is set up appropriately. Future work will extend the junction model along these lines to address issues of learning and nonword processing.

Connectivity and processing

Each localist node was bidirectionally connected to each of 1200 sigmoidal processing units, one unit per bit of orthography (400), phonology (400), and semantics (400). The activation of each localist node was computed by the normalized exponential function,

$$o_j = \frac{e^{\beta I_j + \gamma L_j}}{\sum_k e^{\beta I_k + \gamma L_k}} \quad (1)$$

This function caused the nodes to compete for activation, with o_j being an exponential function of support for word j . Support was summed from two sources, with the strength of each one scaled by the free parameters β and γ . One source was activation of the distributed units, and the other was activation of the neighbouring words. Support from distributed units was given by

$$I_j = \sum_i w_{ji} a_i \quad (2)$$

where w_{ji} was the connection weight from distributed unit i to localist node j , and a_i was the activation of distributed unit i . Each w_{ji} was set equal to +1 or -1 in accordance with the sign of bit i in the pattern for word j . This type of connectivity meant that support increased as the correlation increased between the pattern of activation over the distributed units, and the incoming weight vector.

Support from neighbouring words was given by a novel, Hebbian-like input,

$$L_j = o_j \sum_n o_n \quad (3)$$

where the o_n were activations of all the neighbours of word j . Neighbours of word j were all other words that shared at least 280 out of 400 bits in their orthographic, phonological, or semantic patterns. This neighbourhood support was “self-scaled” (i.e. scaled by o_j) to ensure that support was maximal when activations were evenly distributed in a neighbourhood, and minimal when either o_j or the sum of neighbour activations went to zero. Neighbourhood support provided cooperative interactions among words, and the balance of cooperation and competition was controlled by the ratio of β to γ .

The activation of each distributed unit was computed as a sigmoid bounded by $(-1, +1)$,

$$a_i = \tanh(aI_i + \tau E_i) \quad (4)$$

There were two sources of net input into the distributed units, scaled by the free parameters a and τ . One source came from activations of the localist nodes,

$$I_i = \sum_j w_{ij} o_j \quad (5)$$

where w_{ij} was the connection weight from localist unit j to distributed unit i , and o_j was the activation of localist node j . Each w_{ij} was set equal to $+f_{ij}$ or $-f_{ij}$, where f_{ij} was the log frequency of occurrence of word j in the COALS database, and its sign was set in accordance with the sign of bit i in the pattern for word j . Given that the o_j were normalized to one, each I_i was an average of the w_{ij} , weighted by the o_j .

The other source of input into the distributed units, E_i , was external to the model. It came from environmental inputs such as letters seen or sounds heard. In the current simulations, the only external input was orthographic, in order to simulate the standard tasks of word naming or lexical decision. The E_i were set to -1 or $+1$ in accordance with the orthographic bit pattern of a given input word. The influence of internal and external inputs to the distributed units was balanced by the α and τ free parameters, respectively.

Unit and node activations were computed in discrete time. The distributed units were used to index the model's response to an orthographic input. The model's readiness to respond was measured by the degree to which unit activations were near their asymptotes. This response measure was implemented by the response probability function

$$p(t) = E[(a'_j)^2] \quad (6)$$

where the probability of a response at time step t was equal to the mean square activation value across all phonological units. The phonological units constituted a generic response measure that could be used to simulate both naming response times and lexical decision response times. These tasks differ in important ways, but this simplified response measure was adequate for the pilot model.

The probability of a response over time is illustrated in Figure 3.7 for three example word inputs. The probability response function given by equation (6) allowed for these response distributions to be calculated after only a single presentation of each word to the model. The graphs show that response distributions for individual items are positively skewed, and that items with faster response times also have less variation in their response times. These distributional characteristics were a natural consequence of the model architecture, and they are consistent with the distributional characteristics of real response times to words.

Model tests

The model's free parameters were fit against the mean naming response times for 30,894 words that were collected as part of the Elexicon database (Balota

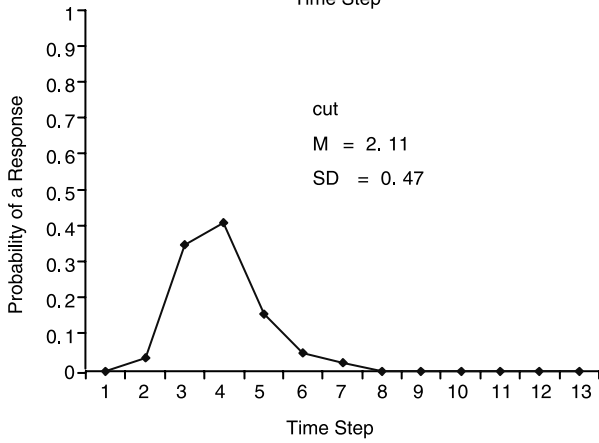
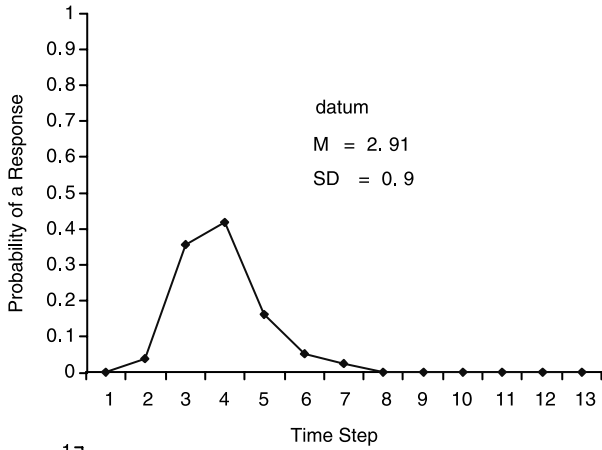
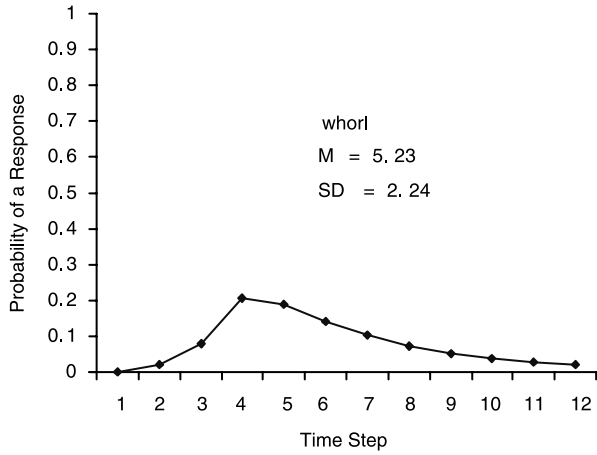


Figure 3.7 Probability density functions of the model response times for three individual words, with means (M) and standard deviations (SD) shown.

et al., 2002; *Balota & Yap*, this volume). This set comprised the full intersection of the model's lexicon with the data available in the Elexicon database at the time of simulation. Model performance was mostly the same for a wide range of parameter values, indicating that model fit was not crucially dependent on any particular set of values. It is well known that the articulatory characteristics of a naming response have a large effect on naming response times (Kessler, Treiman, & Mullennix, 2002; Rastle & Davis, 2002). It is also well known that the length of a word has large effects on naming response times and lexical decision response times (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004). The current model cannot be held accountable for these effects because it does not have a mouth or eyes. Therefore, the effect of articulatory characteristics of the initial phoneme was partialled out of the Elexicon mean naming response times, and the effect of length was partialled out of the Elexicon mean naming and lexical decision response times. The following analyses were conducted on the resulting response time residuals.

The histogram of normalized model response times is graphed in Figure 3.8, along with the normalized response time residuals for word naming and lexical decision data from the Elexicon database. The graph shows a close fit between the simulated and empirical distributions, particularly with respect to positive skew. It is important to note that there were no parameters to directly control the shape of the model response time distribution. Moreover, the parameter values were not adjusted with respect to the response time distribution; they were instead adjusted to maximize the variance in naming response times explained by the model (see next).

In Figure 3.9, the model response times are compared against the Elexicon response times for naming. The model accounted for 18.4% of the variance in naming residuals, and the parameter values were adjusted to maximize this percentage. By comparison, log frequency accounted for 16.6% of the variance when it was regressed against the response time residuals. To determine

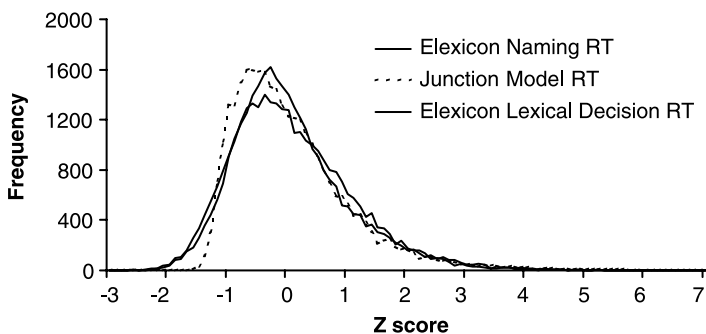


Figure 3.8 Normalized histograms of mean response times (RT) from the junction model, and from mean response time residuals from the naming data and lexical decision data in the Elexicon database.

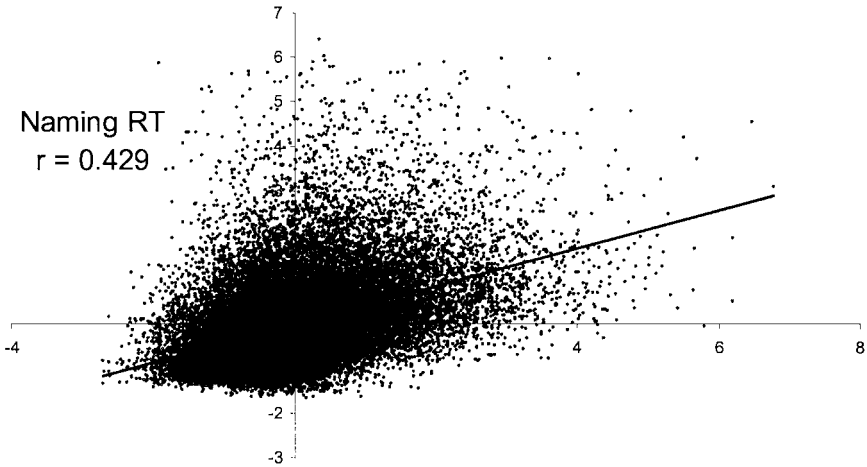


Figure 3.9 Model mean response times (RT) plotted against the naming response time residuals from the Elexicon database, in normalized coordinates.

why the model accounted for this additional variance, log frequency was partialled out of both the model response times and the response time residuals: the model still accounted for 2.6% of the remaining variance. This 2.6% was solely attributable to the competitive and cooperative interactions among words.

In Figure 3.10, the model response times are plotted against the Elexicon response time residuals for lexical decision, in normalized coordinates. The

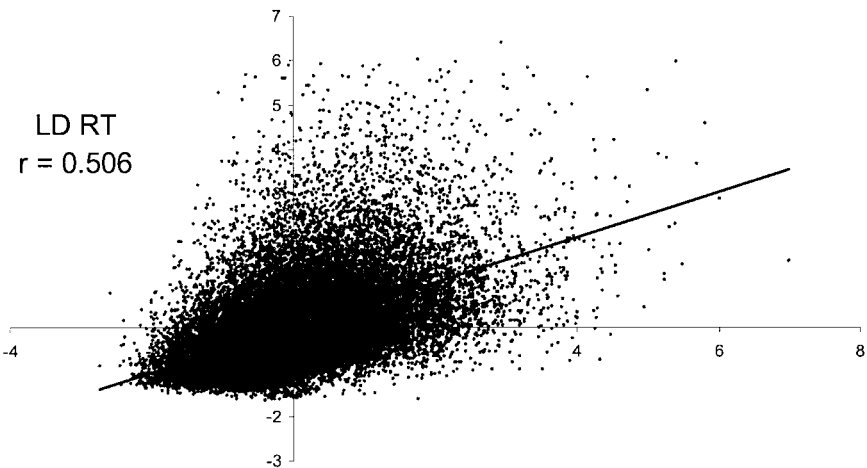


Figure 3.10 Model mean response times plotted against the lexical decision response time (LD RT) residuals from the Elexicon database, in normalized coordinates.

model accounted for 24.7% of the variance in response time residuals, even though the parameter values were not adjusted with respect to these residuals. Log frequency accounted for about the same amount of variance (24.8%). When frequency was removed from the model and from the residuals, the model still accounted for 2.4% of the remaining variance. As with naming response times, this 2.4% was solely attributable to the competitive and cooperative interactions among words.

In addition to mean response times for each item, the Elexicon database provides the standard deviations of response times. As shown in Figure 3.5, the model naturally predicts both the means and standard deviations of item response times. The model's predictions were tested against the database. The model accounted for 7.8% of the item variance in standard deviations for naming, and 5.2% for lexical decision. By comparison, log frequency accounted for 8.2% and 5.9%, respectively. The slightly higher percentages for log frequency mean that the model's ability to account for item standard deviations was primarily due to the way that word frequency had its effect on processing.

Finally, the large-scale junction model was also compared against the two currently dominant models of lexical processing, the triangle (Plaut et al., 1996) and DRC (Coltheart et al., 2001) models of lexical processing. The PMSP and DRC response times were subjected to the same analyses as the junction model response times, and comparisons are shown in Table 3.1. The junction model clearly outperformed the other two models (PMSP simulation 4 performed so well because word frequency was more directly reflected in performance than with the other PMSP simulations). The DRC comparison is somewhat fair because both models are engineered, and the DRC model has 31 free parameters, which are 26 more than the junction model. The PMSP comparison is less fair because the mappings between orthography and phonology were learned in the PMSP model, but not in the junction model. In making either comparison, one must bear in mind that nonword processing was not addressed in the current implementation of the junction model. This will be an important point of comparison in future work that will address nonword processing by incorporating the most recent sequence encoders that support good generalization.

Table 3.1 Proportions of variance in naming response times accounted for by the junction model, compared with the dual-route cascaded (DRC) and Plaut et al. (1996) PMSP models

<i>DRC comparison</i> <i>n = 5190</i>		<i>PMSP comparisons</i> <i>n = 2808</i>				
Junction	DRC	Junction	Sim 1	Sim 2	Sim 3	Sim 4
12.2%	5.1%	14.7%	5.2%	4.1%	2.1%	11.9%

Sim: simulation.

General discussion

It is time to revisit the questions that started off the chapter. First, should we accept the premise that skilled word reading is supported by lexical and sublexical routes of processing? As it stands, the junction model is computational evidence that we should not accept this premise, at least not without more distinguishing evidence. Second, how can we model the learning of orthographic and phonological representations for multisyllabic words? The sequence encoder provides a way, and in doing so, it opens the door to very large-scale models of lexical processing that simulate performance for both mono- and multisyllabic items. The perennial need for such models was heightened when Spieler and Balota (1997) made the well-taken point that item-level performance is an important source of evidence for theories of lexical processing. They showed that the PMSP and DRC models accounted for little variance in performance measures for monosyllabic items, not to mention the fact that they do not simulate performance on multisyllabic items. Balota and his colleagues then upped the ante by providing performance measures for over 30,000 words in the Elexicon database.

The junction model implemented here represents a significant step toward a very large-scale model of lexical processing. The results were encouraging, but more steps lie ahead. Most immediately, the model needs to be extended to address issues of learning and nonword processing, as discussed earlier. These extensions are necessary precursors to embarking on the task of addressing the wide range of results that exist in the literature on lexical processing. For instance, the model will need to be tested on a number of hallmark findings from word naming and lexical decision experiments. It will also need to be tested against a variety of findings from priming and blocking experiments (see *Davis & Kim*, this volume; *Balota & Yap*, this volume). And crucially, it will need to be tested against the data on lexical acquisition and impairments (see *Castles & Nation*, this volume).

These tasks are daunting but doable. The pilot model has already shown that lexical-decision and word-naming data can be simulated. Previous work on the simulation of priming in connectionist models, both weight-based and activation-based, could be incorporated into the junction model. With regard to impairments, performance errors may be generated from noise or damage to units or connections, as in all connectionist models. They may also come from aberrant settings chosen on the basis of specific hypotheses about control parameters of the lexical system. For instance, competition among the localist nodes could be increased by increasing the β parameter, which is equivalent to the input gain parameter in the localist simulations reported by Kello et al. (2005a). Those simulations indicated that high levels of input gain should cause the large-scale model to exhibit a selective impairment in nonword reading, akin to phonological dyslexia. Low levels of input gain should cause a selective impairment in exception word reading.

With regard to acquisition, the localist nodes would need to be formed,

and their weight vectors learned, in coordination with the learning of junction representations. This coordination could be implemented by running two learning algorithms that interact with each other. For instance, back propagation in the sequence encoder could be used to learn junction representations, and the developing junctions could be fed as inputs to a localist learning algorithm. In turn, the localist nodes could bias the mapping that is learned by the sequence encoder. All of these ideas are specific enough to be implemented and tested against empirical data.

In closing, I would like to indulge in a bit of pontification. Computational models of language and cognition serve a number of interrelated purposes (see *Seidenberg & Plaut*, this volume). Broadly speaking, the principle of junctions provides a way to think about cognitive representations that emerge from the demands of perception and action. The principle of junctions is made more tangible by theoretical constructs such as distributed coding and pattern formation in the confluence of competitive and cooperative interactions. These constructs come from the connectionist and complex systems frameworks, which carry with them formalisms for computational modelling. The models are implementations of the constructs, and thereby provide the most tangible incarnation of the principles.

Computational models allow for rigorous tests of the constructs and principles, but they also serve as a wellspring for new ideas. For instance, it was the computational work with input gain that led to the idea that junction representations could be mediated by localist nodes, and still provide a basis for explaining nonword generalization and acquired dyslexia. The use of localist nodes distinguishes the junction model from the PDP “triangle” models at the level of construct, but these models share the principles of distributed coding. Moreover, it was work on the sequencer architecture that allowed us to apply the construct of distributed coding to represent multisyllabic words. The interplay of principles, constructs, and models has driven and will continue to drive progress on the junction model of lexical processing.

Acknowledgements

The author thanks Sally Andrews and Kathy Rastle for comments on an earlier draft of this chapter. The work was funded by awards from NIH and NSF, and does not reflect the views of NSF. Correspondence should be addressed to Christopher Kello, Department of Psychology, George Mason University, 4400 University Drive, Fairfax, VA 22030, USA, ckello@gmu.edu.

References

Balota, D. A., Cortese, M. J., Hutchison, K. A., Neely, J. H., Nelson, D., Simpson, G. B., et al. (2002). *The English Lexicon Project: A web-based repository of descriptive*

- and behavioral measures for 40,481 English words and nonwords. <http://lexicon.wustl.edu>: Washington University. Accessed December 2005.
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283–316.
- Behrmann, M., & Bub, D. (1992). Surface dyslexia and dysgraphia: Dual routes, single lexicon. *Cognitive Neuropsychology*, *9*, 209–251.
- Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*, 221–268.
- Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, *1*, 372–381.
- Coltheart, M. (1996). Phonological dyslexia: Past and future issues. *Cognitive Neuropsychology*, *13*, 749–762.
- Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*, 589–608.
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–256.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.
- Fellous, J.-M., & Linster, C. (1998). Computational models of neuromodulation. *Neural Computation*, *10*, 771–805.
- Funnell, E. (1983). Phonological processes in reading—new evidence from acquired dyslexia. *British Journal of Psychology*, *74*, 159–180.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.
- Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 674–691.
- Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, *87*, 1–51.
- Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720.
- Jordan, M. I. (1986). *Serial order: A parallel distributed processing approach* (No. 8604 ICS Technical Report): University of California at San Diego, La Jolla, CA.
- Kello, C. T. (2003). The emergence of a double dissociation in the modulation of a single control parameter in a nonlinear dynamical system. *Cortex*, *39*, 132–134.
- Kello, C. T. (2004). Control over the time course of cognition in the tempo-naming task. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 942–955.
- Kello, C. T., & Plaut, D. C. (2000). Strategic control in word reading: Evidence from speeded responding in the tempo-naming task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 719–750.
- Kello, C. T., & Plaut, D. C. (2003). Strategic control over rate of processing in word reading: A computational investigation. *Journal of Memory and Language*, *48*, 207–232.
- Kello, C. T., Sibley, D. E., & Colombi, A. (2004). Using simple recurrent networks to

- learn fixed-length representations of variable-length strings. In *Proceedings of the AAAI Symposium on Compositional Connectionism*. Washington, DC.
- Kello, C. T., Sibley, D. E., Plaut, D., & Elman, J. L. (2005b). *Learning representations of linguistic sequences*. Manuscript submitted for publication.
- Kello, C. T., Sibley, D. E., & Plaut, D. (2005a). Dissociations in performance on novel versus irregular items: Single-route demonstrations with input gain in localist and distributed models. *Cognitive Science*, *29*, 627–654.
- Kessler, B., Treiman, R., & Mullennix, J. (2002). Phonetic biases in voice key response time measurements. *Journal of Memory and Language*, *47*, 145–171.
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.
- Monsell, S., Patterson, K. E., Graham, A., Hughes, C. H., & Milroy, R. (1992). Lexical and sublexical translation of spelling to sound: Strategic anticipation of lexical status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 452–467.
- Patterson, K., & Hodges, J. R. (1992). Deterioration of word meaning: Implications for reading. *Neuropsychologia*, *30*, 1025–1040.
- Patterson, K., & Lambon-Ralph, M. A. (1999). Selective disorders of reading? *Current Opinion in Neurobiology*, *9*, 235–239.
- Patterson, K., & Marcel, A. (1992). Phonological alexia or phonological alexia? In J. Alegria, D. Holender, J. Morais, & M. Radeau (Eds.), *Analytic approaches to human cognition* (pp. 259–274). Amsterdam: Elsevier.
- Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193.
- Plaut, D. C., & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, *15*, 445–485.
- Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney (Ed.), *The emergence of language* (pp. 381–415). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- 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*, 56–115.
- Rastle, K., & Davis, M. H. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 307–314.
- Rohde, D. L. T., Gonnerman, L., & Plaut, D. C. (in preparation, 2006). *An improved model of semantic similarity based on lexical co-occurrence*.
- Rosenblatt, F. (1961). *Principles of neurodynamics: Perceptions and the theory of brain mechanisms*. Washington, DC: Spartan Books.
- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.
- Sibley, D. E., & Kello, C. T. (2004). Dissociations between regularities and irregularities in language processing: Computational demonstrations without separable processing components. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings*

- of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 1249–1254). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science*, 8, 411–416.
- Van Orden, G. C., Bosman, A. M. T., Goldinger, S. D., & Farrar, W. T. I. V. (1997). A recurrent-network account of reading, spelling, and dyslexia. In J. W. Donahoe & V. P. Dorsel (Eds.), *Neural-network models of cognition: Biobehavioral foundations. Advances in psychology* (pp. 522–538). Amsterdam: North-Holland/Elsevier.
- von der Malsburg, C. (1981). The correlation theory of brain function. *MPI Biophysical Chemistry, Internal Report*, 81–2. Reprinted in E. Domany, J. L. van Hemmen, & K. Schulten (Eds.), *Models of neural networks II* (1994). Berlin: Springer.
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review*, 9, 625–636.
- Zevin, J. D., & Balota, D. A. (2000). Priming and attentional control of lexical and sublexical pathways during naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 121–135.