Models of visual word recognition

Dennis Norris

Medical Research Council Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 7EF, UK

Reading is a complex process that draws on a remarkable number of diverse perceptual and cognitive processes. In this review, I provide an overview of computational models of reading, focussing on models of visual word recognition—how we recognise individual words. Early computational models had ‘toy’ lexicons, could simulate only a narrow range of phenomena, and frequently had fundamental limitations, such as being able to handle only four-letter words. The most recent models can use realistic lexicons, can simulate data from a range of tasks, and can process words of different lengths. These models are the driving force behind much of the empirical work on reading. I discuss how the data have guided model development and, importantly, I also provide guidelines to help interpret and evaluate the contribution the models make to our understanding of how we read.

From boxes and arrows to computational models of reading

Reading is an impressive human achievement that requires coordinated mastery of a constellation of perceptual and cognitive processes ranging from low-level visual perception to recognition of word forms, phonological processing, eye-movement control, and all of the higher-level linguistic processes required to recover the meaning of the written words. Understanding each of these processes is hard but understanding how they operate as a whole presents an even greater challenge. Early models of reading were predominantly of the ‘box-and-arrow’ type. However, even the most influential of these models – Morton’s logogen model [1] – had very little to say about exactly what went on in the boxes or what information flowed along the arrows. The situation changed dramatically with the development of computational models of reading in the early 1980s. These models made clear statements about what was supposed to be going on in the boxes and we could now work out exactly what the models predicted. These first models were simple connectionist networks (see Glossary). Since then, models have increased in their ability to produce ever-more-accurate simulations of an increasingly wide range of challenging data. New models continue to emerge, with several of the more recent models departing from the connectionist tradition. This review concentrates primarily on more recent models that try to explain the core process that uniquely characterises reading: recognising words as visual objects. These visual objects can then make contact with the full range of representations in the reader’s mental lexicon.

Why ‘computational’ models?

Models of reading are almost invariably computational models. This is true of theories of word identification [2–11], reading aloud [12–15], morphology [16], and eye movements in reading text [17–19] and of models of spoken word recognition [20,21]. How has the field come to be so reliant on computational models? After all, in many cases the underlying principles of the models are simple. However,

Glossary

Bayes theorem: a mathematical procedure for updating probabilities or beliefs in the light of new evidence. In the case of word recognition, the probability of a word given the input, or evidence, is as follows:

\[ P(\text{word}|\text{evidence}) = \frac{p(\text{word}) \times p(\text{evidence}|\text{word})}{\sum_{i} p(\text{word}_i) \times p(\text{evidence}|\text{word}_i)} \]

Connectionism: models expressed as artificial neural networks; this includes, for example, the IA model. These models are intended to capture general properties of neurons, or neuronal populations.

Interactive activation (IA) model: the first, and still most influential, form of connectionist model of word recognition. Words are represented as nodes in a network that are connected by inhibitory links (see Figure 1 in main text).

Lexical competition: in both IA models and Bayesian models, neighbouring words compete with each other for recognition. In IA models, this is due to the inhibitory connections between word nodes.

Lexical decision: the most common laboratory task for studying word recognition. Participants are required to decide whether a string of letters is a word or not (a nonword).

Masked priming: a variant on the lexical decision task in which the target is preceded by a briefly presented prime, which can be a word or a nonword. Participants are rarely aware of the prime. The prime is usually presented in lower case and the target in upper case to minimise physical overlap. Masked priming is most commonly used to address questions about the representation of orthography.

Neighbourhood density: a measure of how similar a word is to other words. A common measure is Coltheart’s N [34]: how many other words can be formed by changing a single letter in a word? According to this definition, only words of the same length can be neighbours. A more flexible measure is given by a Levenshtein distance metric. This measures similarity in terms of the number of ‘edits’ – insertions, deletion, and substitutions – so WORD and WORDS will now be considered to be neighbours. The OLD20 is the average distance of the 20 closest neighbours.

Open bigrams: a proposal that the order of letters in a word is coded in terms of a set of ordered letter pairs, which may be non-contiguous. WORD might be coded as WO, WR WD, OR, OD, or RD

Reaction time (RT) distribution: RTs in tasks like lexical decision are generally positively skewed. Variables like word frequency rarely shift only the mean of the distribution, but usually the form of the distribution, too. Accounting for these changes is a challenge for computational models.

Word-frequency effect: by far the strongest influence on how readily a word can be identified is its frequency of occurrence in the language; words that occur very often in the language are recognised more quickly than low-frequency words. The speed and ease with which words can be recognised is an approximately logarithmic function of word frequency.
even with a deep understanding of the principles and mathematical foundations underlying the models, it is almost impossible for theorists to be sure how their models will behave. The reason is straightforward: the behaviour of the models is not determined simply by the high-level theoretical principles themselves, but emerges as an interaction between those principles and the contents of the lexicon. How any one word will be processed depends critically on the nature of the other words in the lexicon. Given that some of the models now use lexicons containing many tens of thousands of words, the only way to be sure exactly what the theories predict is to implement them as computational models. However, although there is universal agreement that computational models are to be preferred over older verbal or box-and-arrow models (logogen), there is a continuing debate about the most useful style of model.

Modelling style

The earliest and most influential style of computational model is the interactive activation (IA) model [11,22] (Figure 1) – one of the first connectionist or ‘neural-network’ cognitive models. In almost all IA models, letter features, letters, and words are represented as nodes in a network (a ‘localist’ representation). IA networks generally have no capacity to learn. Although IA models remain popular (the Spatial Coding Model [2] and the dual-route cascaded (DRC) model [12] are two recent examples in this tradition), many connectionist models incorporate learning mechanisms and use ‘distributed representations’. This is most common in models of reading aloud [15,23–25]. In models using distributed representations, words are usually not represented by a single node, but as a pattern of activations over a set of nodes.

Connectionist models are often favoured because they appear ‘brain like’ [26] or ‘neurally inspired’ [27]. An alternative view is that we know so little about how words might be represented in the brain, or how the relevant neural computations are performed, that we should formulate our models at a more abstract level that makes no claims about implementation and concentrate instead on understanding the nature of those computations [28]. Many of these models are therefore expressed primarily in terms of computational procedures or mathematical formulae. Table 1 lists the most influential computational models and indicates which style of modelling they use and the primary phenomena they have been developed to explain. Note that although the primary focus here is on models of visual word recognition, the table lists a broader range of models, including connectionist models of reading aloud and models of eye-movement control in reading.

When comparing different styles of model, appearances can be deceptive. For example, Ratcliff’s drift–diffusion model (DDM) [29,30] is usually expressed mathematically, but can trivially be recast as a simple connectionist network if the network is crafted to compute exactly the right function. That would change how the model looks, but would not alter the underlying theory or explanation. Similarly, Figure 1 contrasts an IA model with Bayes formula, which is the basis of the Bayesian Reader (BR). However, IA models can also be formulated to compute Bayes theorem. Next, I give a brief description of the three most recent models of visual word recognition [2,6,8] that also illustrates contrasting modelling styles. One is a connectionist model and the other two are mathematical/computational.

The Spatial Coding Model (SCM) [2] is based on the IA framework. The original IA model could simulate words of a fixed length only. The SCM has been further developed to enable it to process words of varying lengths and to simulate masked priming. The distinctive feature of the SCM model is the way that it represents the order of letters in the input in terms of an activation gradient over letter positions (Figure 2). The model incorporates a matching rule that is relatively insensitive to exactly where words begin in the input (TOP will be activated in STOP) and also tolerates minor changes in the relative position of letters (JUDGE will activate JUDGE; see Box 2).

The Letters in Time and Retinotopic Space (LTRS) [8] model was developed primarily to account for data on perceptual identification and masked priming. It assumes that information about letter identity and letter order accumulates stochastically over time. Importantly, although there is variability in the time at which a letter is identified, letters and their associated order information are always identified correctly. Given the prime JUDGE in a masked priming task, there is some probability that, at the end of the prime, the only evidence that has accumulated might be JU*n*GE, where *n* corresponds to one or more unknown letters. This would be consistent with the word JUDGE and produce priming. However, given the prime JUNPE, if either N or P are identified this will be inconsistent with the target and not produce priming (Box 2).

The LTRS model makes no specific assumptions about the precise form of representations – any representation
Table 1. Major computational models of reading organised in terms of their primary focus.a,b

| Model                       | Style                | Task         | Phenomena                           | Large lexicon |
|-----------------------------|----------------------|--------------|-------------------------------------|---------------|
| Models of visual word recognition |                      |              |                                     |               |
| IA [11,22]                  | IA                   | PI           | Word-superiority effect             |               |
| Multiple read-out [3]       | IA                   | PI, LD       | Word-superiority effect             |               |
| SCM [2]                     | IA                   | LD, MP       | Letter order                        |               |
| BR [4–6]                    | Math/comp            | LD, MP       | Word frequency, letter order, RT distribution | √             |
| LTRS [8]                    | Math/comp            | MP, PI       | Letter order                        |               |
| Overlap [66]                | Math/comp            | PI           | Letter order                        |               |
| Diffusion model [30]        | Math/comp            | LD           | RT distribution, word frequency     |               |
| SERIOL [7]                  | Math/comp            | LD, MP       | Letter order                        |               |
| Models of reading aloud     |                      |              |                                     |               |
| CDP++ [13]                  | Localist/symbolic    | RA           | Reading aloud                       | √             |
| DRC [12]                    | IA                   | RA, LD       | Reading aloud                       |               |
| Triangle [24,25]            | Distributed connectionist | RA   | Reading aloud                       |               |
| Sequence encoder [15]       | Distributed connectionist | RA  | Reading aloud                       | √             |
| Junction model [50]         | Distributed connectionist | RA  | Reading aloud                       | √             |
| Models of eye-movement control in reading |                      |              |                                     |               |
| E-Z reader [17,18]          | Symbolic             | R            | Eye movements                       |               |
| SWIFT [19]                  | Symbolic             | R            | Eye movements                       |               |
| Model of morphology         |                      |              |                                     |               |
| Amorphous discriminative learning [16] | Symbolic network | Self-paced reading, LD | Morphology | √             |

The table also indicates the modelling style or framework, the main task that the model simulates, the main phenomena that the model simulates (not exhaustive), and whether the model uses a realistically sized lexicon. Note that the review concentrates on ‘Models of visual word recognition’.

Abbreviations: Math/comp, mathematical or computational; LD, lexical decision; PI, perceptual identification; RA, reading aloud; MP, masked priming; R, natural reading.

will do, providing it has the correct set of properties. As Adelman notes [8], several representations would satisfy the requirements – for example, a representation involving letters and open bigrams – but the specific choice of representation makes no difference to the model predictions. Indeed, an important contribution of LTRS is to show how a wide range of priming data can be explained while making few assumptions about the exact form of representations.

As with the LTRS model, the BR [6] is formulated at an abstract level that makes no assumptions about implementation and as few assumptions as possible about representations. The aim of the BR is to see how much can be explained simply by assuming that readers make near-optimal decisions based on the accumulation of noisy evidence. The model is optimal in the sense that, for a given level of accuracy, it will identify words based on the fewest number of samples possible; that is, as fast as possible [31]. Bayes theorem provides the optimal procedure for combining uncertain evidence with knowledge of prior probability.

In the model, letters are represented as vectors describing coordinates in a multidimensional space. The dimensions could be considered to correspond to letter features, although they also encode positional information. At each time step, the model accumulates a noisy sample from the input that is created by adding noise to the input vector. As more samples are accumulated, the model’s estimate of the true value of the input becomes more precise and hence the identity and position of the letters is known with greater certainty. For each word in the lexicon, the model computes the likelihood of observing the input, given that the word would have generated that input \( P(\text{evidence}|\text{word}) \). The model also knows the frequency of each word \( P(\text{word}) \). From this, it can use Bayes theorem (Figure 1) to compute...
the probability of each word given the input \( P(\text{word}|\text{evidence}) \). A word can be identified when this probability exceeds some predetermined threshold. In a connectionist model, all words will have some degree of activation. In the BR, activation becomes something much more specific: probability. To focus on the core principles, the model incorporates many simplifying assumptions about the nature of the visual information available. For example, all letters in a word are assumed to be equally perceptible.

In the BR the focus is on optimal decision making. This means that the model naturally accounts for differences between tasks such as lexical decision, perceptual identification, and masked priming. Different tasks require different decisions; therefore, the optimal decision process must necessarily be different, too. Additionally, a Bayesian model must necessarily take account of prior probability, which gives a natural explanation for the word frequency effect. As already noted, the way any one word is processed depends on its relation to other words in the lexicon. The way that words influence each other is generally viewed as a process of lexical competition.

**Lexical competition**

To recognise a word, the reader must accumulate enough evidence to distinguish that word from perceptually similar words: their lexical neighbours. Perceptually similar words must compete with each other for recognition. All current models incorporate some form of lexical competition, although the way that competitive process operates can appear to be different in models that produce very similar behaviour (Box 1 and Figure 1). They also incorporate different assumptions about the form of the perceptual and orthographic representations of words and the way they are processed. Words that are considered to be close neighbours in one model might not be in another [32,33]. This is most apparent in the way different models make contrasting assumptions about the way letter order is represented (Box 2).

Almost all studies of neighbourhood effects use Coltheart’s N [34] as a measure of neighbourhood density. This metric considers only words of equal length to be neighbours. However, words of different lengths such as ‘hat’ in ‘that’ also act as competitors [35–38]. A better measure of density that accounts for more unique variance in lexical decision times is provided by the orthographic Levenshtein distance (OLD20) [39]. The Levenshtein edit distance is given by the number of edits (insertions, deletions, and substitutions) required to transform one word into another. The OLD20 is based on the average edit distance of the 20 nearest neighbours.

**Laboratory tasks**

Although the goal of models of word recognition is to understand normal reading, the only overt behaviour that readers generally produce is to move their eyes. Eye-movement data are hugely informative, but it is rarely practical to collect large amounts of data using carefully controlled stimuli. Many researchers therefore turn to more tractable laboratory tasks such as lexical decision, word naming, and masked priming. This leads to two distinct modelling enterprises. Whereas models of eye-movement control during reading tend to make simplifying assumptions about how individual words are identified [17–19], models of word recognition rarely consider how they might be integrated with models of reading. The use of laboratory tasks poses an additional modelling challenge. Although it is tempting to think of tasks like lexical decision as being direct measures of the time taken to identify a word, each of the tasks engages some additional task-specific processing. For the models to fit the data, they must simulate task performance as well as word identification itself.

Fortunately, the results tend to be similar in research in which the same phenomena have been studied using eye movements and lexical decision [40–42]. However, there is one area where different tasks do produce different results. As noted above, all current models incorporate some form of lexical competition; words with many neighbours should therefore be recognised more slowly than words with few neighbours because they suffer from more competition [38,43]. However, in lexical decision this pattern is reversed [39]. In IA models [2,3,12], this finding is an embarrassment because the networks just have to predict that recognition will be slowed by competition. To overcome this problem, the models have to be modified by adding a decision process that is sensitive to the overall activation in the lexicon [3,12]. More neighbours produce more overall

---

**Box 1. Styles of modelling: IA models versus Bayesian theories**

IA models (see Figure 1 in main text) have several appealing features. One is that they are relatively easy to understand. The basic principle is one of competition between word nodes. Words receive activation in proportion to how well they match the input and nodes compete with each other by means of inhibitory connections. The best-matching word will win the competition but be slowed down by competition from similar words. The most advanced IA model is the SCM [2]. The SCM differs from earlier IA models in that it can deal with words of different lengths. This allows it to simulate a far wider range of phenomena than earlier models. One concern with IA models is that the networks generally require many parameters whose exact values have no principled motivation. For example, how much inhibition should there be between words or how should the models implement the effect of word frequency? In a Bayesian model [6,9,12], such questions do not arise; the precise treatment of lexical competition and word frequency follows automatically from the theoretical claim that readers approximate ideal Bayesian decision makers. Figure 1, in main text, shows a simple IA model and Bayes theorem. Although a connectionist network and an equation look like very different things, they achieve similar ends. Each word node in the IA model sums its perceptual input from letter or feature nodes. Because each word node is connected to every other node, all nodes receive the same amount of inhibition, where that inhibition is proportional to the sum of all other nodes. According to Bayes theorem, the probability of each word is a function of the evidence for that word (called the likelihood) divided by the evidence for all other words. There is a clear parallel between the two formalisms. Consequently, a properly configured network could compute Bayes theorem, but it could also compute a range of different functions. Would we gain anything by implementing a Bayesian model as a connectionist network? A network implementation would simply compute exactly the same function and produce exactly the same simulations, but it would make it harder to appreciate the importance of the theoretical claim that readers were approximating ideal Bayesian decision makers. As Anderson [28] noted, ‘if two theorists propose two sets of mechanisms in two architectures that compute the same function, then they are proposing the same theory’.
 activation which leads to faster responses. They can then account for the opposite pattern of data from the one they naturally predict.

This situation highlights an important contrast between theories that begin by postulating a particular mechanism [2,11,22] and those that focus instead on higher-level computational principles [4–6]. For IA models, the problem is that their explanation of word recognition lies in the details of the mechanism. If the behaviour changes, the mechanism must somehow be changed, too.

In the BR [4–6], the optimal decision must necessarily differ between different tasks. In a perceptual identification task, participants need to select one word from among all the words in the lexicon. In lexical decision, the participant’s task is not to select a single word, but to press a button when they are confident that the input is a word rather than a nonword. The optimal way to respond is to pool the evidence over all words that are similar to the input [4]. Words in dense neighbourhoods will therefore be responded to faster in lexical decision. The BR has to predict that neighbourhood effects will vary depending on the task. They should be facilitatory in lexical decision but inhibitory in tasks requiring identification of a unique word. This follows directly from the idea that readers approximate ideal Bayesian decision makers. By contrast, IA models naturally produce inhibition. To fit the data they can be modified to produce facilitation, but they do not explain why different tasks should produce different results.

The rise of the megastudy

Until relatively recently, the standard recipe for a study of reading would be to carefully select small subsets of words that varied on one or two measures of interest and then to compare them using either a lexical decision task or a speeded naming task. However, we now have access to several large-scale databases, or megastudies, containing lexical decision data for between ten and 40,000 words. The largest of these, the English Lexicon Project (ELP) [44], contains 4 million word-recognition trails collected from over 1200 participants. Data for the ELP was collected in the USA, but there are now similar databases for British English [45], Dutch [46], and French [47]. The ELP also contains data on naming as well as lexical decision. Eye movement data is available from the Dundee corpus [48], which was derived from ten English and ten French participants each reading about 50,000 words. Many hypotheses can therefore be tested by performing virtual experiments on the databases.

Keuleers et al. [45] performed several such experiments where they compared item reaction times (RTs) from previous experiments on word frequency, regularity, feedforward consistency, age of acquisition, polysemy, and neighbourhood density with corresponding item RTs in the British Lexicon Project (BLP). In some cases, the BLP data did not show the same effects as in the original studies. Perhaps the theoretically most significant ‘failure to replicate’ in these virtual experiments was that the BLP did not consistently reveal a facilitatory effect of neighbourhood density (Coltheart’s N). However, the BLP does show the expected correlation with measures of neighbourhood density, albeit a slightly smaller correlation than that seen in the ELP [6]. All of the megastudies show a similar pattern. Yap and Balota [49] presented an analysis of both lexical decision and naming latencies of 6115 monomorphic multisyllabic words from the ELP. They examined the influence of a range of measures including word frequency, letter and syllable length, phonological and orthographic neighbourhood density, and spelling-to-sound consistency. These factors accounted for about 61% of the total variance in both naming and lexical decision.

The rise of the megastudy has raised the bar in terms of what we expect from our computational models. Why stop at just being able to simulate the effect of, say, spelling-to-sound regularity or neighbourhood density using a small set of carefully controlled stimuli? Now we can ask how well the models can simulate item-level RTs for all of the words in the databases. Modellers have started to rise to this challenge. Yap and Balota [49] analysed simulated data from Kello’s [50] junction model and Perry et al.’s [13] Connectionist Dual Process (CDP++) model (see Table 1 for more information on these models) in the same way that they had analysed the human data. With some exceptions, they found that both models were sensitive to the same factors as were human readers. The CDP++ model [13] of reading aloud has been used to simulate reaction times for over 32,000 words, 17,841 of which were in the ELP. The BR [6] simulates lexical decision times for over 26,000 words from the ELP and most of the words in the British, Dutch, and French lexicon projects. Other models simulate smaller but still substantial portions of the megastudy

Box 2. The representation of letter order

As witnessed by the ease with which we can read the famous ‘Cmabridge Uinervtstv’ email (http://www.mrc-bcu.cam.ac.uk/people/dennis.norris/personal/cambridgeemail), readers are remarkably tolerant of changes in the order of letters in a word [70,71]. For example, in the masked priming task, a prime constructed by transposing two letters of the target word produces as much priming as an identity prime (jugde-JUDGE versus judge-JUDGE) and much more than a prime where the same two letters are changed (junpe-JUDGE) [72–75]. This excludes the simplest possible theory of letter coding in which letters have position-specific codes. Under that scheme, a ‘d’ in position 4 is a completely different entity from a ‘d’ in position 3, so jugde should produce no more priming than junpe. Figure 2, in main text, illustrates three alternative letter-coding schemes. Open Bigram coding appeals to a form of local-context coding using pairs of letters [32,76–78]. JUDGE and jugde are deemed to be very similar because they share nine of ten open bigrams, whereas JUDGE and junpe share only three. In models using noisy coding of position or order [6,66,68,79,80], letter order is simply represented as a sequence of letters, as would be found in a dictionary. JUDGE and jugde are similar because uncertainty over the exact position of the letters means that the noisy perceptual input generated by jugde could have been produced by judge. In the SCM, order is represented as a gradient of activation over letter nodes [2]. Studies have shown that priming can also be produced when letters from the target are deleted or other letters are inserted [75,78,81–83]. There are now three computational models that can generate very accurate simulations of most of the experimental data [2,6,8], but none of these relies on open bigrams. All of these models appeal to some form of noisy sampling or noisy coding. One major challenge for open bigram models is to explain how it is that fo primes OF [84]; given that the prime and target do not have any open bigrams in common, there should be no priming. By contrast, this result is exactly what would be expected from all of the other models.
items [2,51,52]. Although the megastudies are an invaluable resource, they have limitations. Whether for lexical decision or for reading aloud, the correlations between the megastudies, or earlier smaller-scale studies, never exceed 0.7 [13,45]. This is not greatly different from the split-half correlations in the BLP [45]. Lexical decision and naming data are fundamentally noisy [53]. Even the same subjects will respond differently on different occasions [54,55]. The studies use different equipment and different nonwords and even vary regarding whether words are presented in upper case (ELP) or lower case (the British, Dutch, and French lexicon projects). Even more importantly, they use different participants with different linguistic experience. The most obvious consequence of this variability between megastudies is that there is an upper limit on how much variance we can expect models to account for. Current models can achieve correlations of about 0.6 with human RTs. Given that the maximum correlation between megastudies is only about 0.7, it might appear that there is only limited room for improvement. However, this does not mean that the models are so good that they cannot be developed further. For example, currently none of the models has the ability to simultaneously model orthographic, phonological, and semantic effects.

The megastudies confirm that the single most powerful determinant of lexical decision or naming speed is the logarithm of the word’s frequency of occurrence in the language (although there is debate about the exact form of this function [56,57]). So how do the models explain the word-frequency effect? Most models simply build the effect in without offering any explanation of why things should be that way. For example, connectionist learning models almost always present words during training in proportion to the logarithm of their frequency, not their actual frequency [13,50]. However, a Bayesian model must take account of prior probability (Figure 1); that is, its frequency. When this is combined with the assumption that perception involves the accumulation of noisy evidence, this automatically produces the observed logarithmic relation between frequency and RT [4]. That is, the Bayesian model delivers the log frequency function for free and this explains why we should observe a logarithmic function rather than any other.

### Beyond mean RT: simulating variability

The usual target for models of word recognition is mean RT. However, even more information can be extracted from the data by examining the distribution of RTs and how they change as a function of stimulus type [58–60,30,61,62] or participant group [63,64]. IA models therefore always respond to the same word in exactly the same way and in exactly the same amount of time. This means that they are unable to simulate RT distributions (although see [65]).

By contrast, evidence-accumulation models [4–6,8,30,66–68] start from the assumption that perception is a fundamentally noisy process and that the task of the perceptual system is to make the best use of that noisy information. The most successful of these models is Ratcliff’s DDM [29,69], which is usually applied to two-choice RTs and can therefore be used to model RT distributions in lexical decision. In the DDM, evidence is accumulated as a sequence of noisy samples until the total evidence reaches a ‘yes’ or ‘no’ decision boundary. The DDM gives a very accurate fit to a range of lexical decision data [30] and provides some interesting insights. For example, it was shown that word frequency influenced the rate at which evidence was accumulated. Norris [5] showed that this pattern follows directly from the BR’s account of word frequency. Note that some of these data can be simulated in an IA model by adding a leaky accumulator decision process to the output [65].

### Box 3. Evaluating models

Given the wide range of computational models available, how should we set about evaluating them? What makes one model better than another? The usual selling point of a model is to emphasise how well it fits the data. A model that cannot fit the data is clearly of little value. However, neither is a model that can fit any pattern of data that might possibly be observed [85]. A partial solution to this problem is to use formal methods for comparing models with different numbers of free parameters [86–88] that penalize models with greater flexibility. However, sometimes flexibility does not come from the settings of free parameters but from ad hoc modifications to the structure of the model designed to accommodate new pieces of awkward data. Of course, in itself, extending and developing models is no bad thing, but models should be ‘nested’ [89,90] such that any new version should still be able to simulate the data covered by the old model. Given that old models evolve and new models need to fit the data to be published, models tend to converge.

Perhaps the most important question to ask of any model is whether it provides a good explanation of the data. A computer program that happened to simulate the data but whose operation was opaque would make little contribution to our understanding of word recognition [91]. The model should be a computational implementation of a theory and the explanation is a property of the underlying theory rather than the model [92]. We need to look beyond the particulars of the model and ask how the principles and assumptions of the theory explain how words are recognised. Ideally, we would also like a theory to shed some light on why our perceptual processes operate in the way they do. Indeed, addressing the why question is one of the main goals of the Bayesian approach [93,94]. Box [95] famously stated that ‘all models are wrong, but some are useful’. A model that is wrong but useful may be better than a model that is ‘right’ (fits the data) but of little use in helping to explain the phenomena of interest. We should value models for their theoretical insights and not just for their ability to fit the data.

### Concluding remarks

Modelling word recognition began with small-scale simulations using perhaps a thousand words, all of the same length [22]. The target for simulation was perceptual identification scores from a few small datasets. Models can now perform large-scale simulations of data from tens of thousands of words. The scope of the models has been expanded to cover tasks like lexical decision, masked priming, reading aloud, and eye-movement control. Models now simulate a far wider range of empirical findings than their predecessors and some can simulate RT distributions as well as means. Although, comparing models is rarely straightforward (Box 3), much of the empirical work in the area is now targeted at testing differential predictions of the models. Despite their successes, current models all have limitations (Box 4). In particular, individual models tend to focus on a single domain of behaviour, such as...
Box 4. Outstanding questions

- Current models each deal only with subcomponents of the reading process. One of the greatest challenges is to produce an integrated model of reading. For example, we have no process models of how morphological or semantic representations interact with orthographic processing.
- Many of the challenges facing models of reading are shared with all models of visual perception. For example, we know little about how readers achieve translational invariance; that is, the ability to recognise words presented in different locations or to recognize morphemes embedded in longer words (e.g., like in dislike). How are ‘a’, ‘A’, and ‘a’ all treated as instances of the letter ‘a’?
- How is word recognition influenced by differences between languages and writing systems?
- Can we make models more relevant to the understanding of reading disorders? Current computational models of word recognition concentrate on fluent reading in the stable adult system and have little to say about how reading develops and how that development might be impaired.

reading aloud, eye movements, or lexical decisions. There is a need for more integrated theories of word recognition.

References

1 Morton, J. (1989) The interaction of information in word recognition. Psychol. Rev. 76, 165–178
2 Davis, C.J. (2010) The spatial coding model of visual word identification. Psychol. Rev. 117, 713–758
3 Grainger, J. and Jacobs, A.M. (1996) Orthographic processing in visual word recognition: a multiple read-out model. Psychol. Rev. 103, 518–565
4 Norris, D. (2006) The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. Psychol. Rev. 113, 327–357
5 Norris, D. (2009) Putting it all together: a unified account of word recognition and reaction-time distributions. Psychol. Rev. 116, 207–219
6 Norris, D. and Kinoshita, S. (2012) Reading through a noisy channel: why there’s nothing special about the perception of orthography. Psychol. Rev. 119, 517–545
7 Whitney, C. and Corneliussen, P. (2008) SERIOL Reading. Lang. Cogn. Process. 23, 143–164
8 Adelman, J.S. (2011) Letters in time and retinotopic space. Psychol. Rev. 118, 570–582
9 Diependaele, K. et al. (2010) Fast phonology and the Bimodal Interactive Activation Model. Eur. J. Cogn. Psychol. 22, 564–778
10 McClelland, J.L. and Rumelhart, D.E. (1981) An interactive activation model of context effects in letter perception: part 1. An account of basic findings. Psychol. Rev. 88, 375–407
11 Rumelhart, D.E. and McClelland, J.L. (1982) An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. Psychol. Rev. 89, 60–94
12 Coltheart, M. et al. (2001) DRC: A dual route cascaded model of visual word recognition and reading aloud. Psychol. Rev. 108, 204–256
13 Perry, C. et al. (2010) Beyond single syllables: large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. Cogn. Psychol. 61, 106–151
14 Perry, C. et al. (2007) Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. Psychol. Rev. 114, 273–315
15 Sibley, D.E. et al. (2008) Large-scale modeling of wordform learning and representation. Cogn. Sci. 32, 741–754
16 Baayen, R.H. et al. (2011) An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. Psychol. Rev. 118, 438–481
17 Reichle, E.D. et al. (2009) Using EZ Reader to model the effects of higher level language processing on eye movements during reading. Psychon. Bull. Rev. 16, 1–21
18 Reichle, E.D. et al. (2012) Using E-Z Reader to simulate eye movements in nonreading tasks: a unified framework for understanding the eye-mind link. Psychol. Rev. 119, 155–185
19 Engbert, R. et al. (2005) SWIFT: a dynamical model of saccade generation during reading. Psychol. Rev. 112, 777–813
20 Norris, D. and McQueen, J.M. (2008) Shortlist B: A Bayesian model of continuous speech recognition. Psychol. Rev. 115, 357–395
21 McClelland, J.L. and Elman, J.L. (1986) The TRACE model of speech perception. Cogn. Psychol. 18, 1–86
22 McClelland, J. and Rumelhart, D. (1981) An interactive activation model of context effects in letter perception: part 1. An account of basic findings. Psychol. Rev. 88, 370–407
23 Seidenberg, M.S. and McClelland, J.L. (1989) A distributed, developmental model of word recognition and naming. Psychol. Rev. 96, 523–568
24 Harm, M.W. and Seidenberg, M.S. (2004) Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. Psychol. Rev. 111, 662–720
25 Plaut, D.C. et al. (1996) Understanding normal and impaired word reading: computational principles in quasi-regular domains. Psychol. Rev. 103, 56–115
26 Clark, A. (1989) Microgenetic: Philosophy, Cognitive Science and Parallel Distributed Processing. MIT Press
27 Rumelhart, D. (1989) The architecture of mind: a connectionist approach. In Foundations of Cognitive Science (Posner, M., ed.), pp. 133–159, MIT Press
28 Anderson, J.R. (1990) Adaptive Character of Thought. Erlbaum
29 Ratcliff, R. (1978) A theory of memory retrieval. Psychol. Rev. 85, 59–109
30 Ratcliff, R. et al. (2004) A diffusion model account of the lexical decision task. Psychol. Rev. 111, 159–182
31 Wald, A. (1947) Sequential Analysis, Wiley
32 Whitney, C. (2008) Comparison of the SERIOL and SOLAR theories of letter-position encoding. Brain Lang. 107, 170–178
33 Davis, C.J. and Bowers, J.S. (2006) Contrasting five different theories of letter position coding: evidence from orthographic similarity effects. J. Exp. Psychol. Hum. Percept. Perform. 32, 535–557
34 Coltheart, M. et al. (1977) Access to the internal lexicon. In Attention and Performance VI (Dornic, S., ed.), pp. 535–555, Erlbaum
35 Bowers, J. et al. (2005) Automatic semantic activation of embedded words: is there a “hat” in “that”? J. Mem. Lang. 52, 131–143
36 Davis, C.J. et al. (2009) Redefining the orthographic neighborhood: the role of addition and deletion neighbors in lexical decision and reading. J. Exp. Psychol. Hum. Percept. Perform. 35, 1550–1570
37 Davis, C.J. and Taft, M. (2005) More words in the neighborhood: interference in lexical decision due to deletion neighbors. Psychon. Bull. Rev. 12, 904–910
38 Weinagartner, K.M. and Rayner, K. (2013) Lexical embeddings produce interference when they are morphologically unrelated to the words in which they are contained: evidence from eye movements. J. Cogn. Psychol. 24, 179–188
39 Yarkoni, T. et al. (2008) Moving beyond Coltheart’s N: a new measure of orthographic similarity. Psychon. Bull. Rev. 15, 971–979
40 Kuperman, V. and Drieghe, D. (2013) How strongly do word reading times and lexical decision times correlate? Combining data from eye movement corpora and megastudies. Q. J. Exp. Psychol. (Hove) 66, 563–580
41 White, S.J. (2006) Eye movement control during reading: effects of word frequency and orthographic familiarity. J. Exp. Psychol. Hum. Percept. Perform. 34, 205–223
42 Smith, N.J. and Levy, R. (2010) Fixation durations in first-pass reading reflect uncertainty about word identity. In Proceedings of the 32nd Annual Meeting of the Cognitive Science Society. Cognitive Science Society
43 Paterson, K.B. et al. (2009) Inhibitory neighbor priming effects in eye movements during reading. Psychon. Bull. Rev. 16, 49–50
44 Baleta, D.A. et al. (2007) The English Lexicon Project. Behav. Res. Methods 39, 445–459
45 Keuleers, E. et al. (2012) The British Lexicon Project: lexical decision data for 28,730 monosyllabic and disyllabic English words. Behav. Res. Methods 44, 287–304
46 Keuleers, E. et al. (2010) Practice effects in large-scale visual word recognition studies: a lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. Front. Psychol. 1, 174
47 Ferrand, L. et al. (2010) The French Lexicon Project: lexical decision data for 38,840 French words and 38,840 pseudowords. Behav. Res. Methods 42, 488–496
Ratcliff, Norris, Dufau, Andrews, Balota, Murray, Adelman, Rey, the cost. Unmasked priming. In From Inmarks to Ideas: Current Issues in Lexical Processing (Andrews, S., ed.), pp. 50–75, Psychology Press

Sibley, D. et al. (2010) Learning orthographic and phonological representations in models of monosyllabic and bisyllabic naming. Eur. J. Cogn. Psychol. 22, 650–668

Whitney, C. (2011) Location, location, location: how it affects the neighborhood (effect). Brain Lang. 118, 90–104

Rey, A. et al. (2009) Item performance in visual word recognition. Psychon. Bull. Rev. 16, 600–608

Adelman, J.S. et al. (2013) The unexplained nature of reading. J. Exp. Psychol. Learn. Mem. Cogn. 39, 1037–1053

Diependaele, K. et al. (2012) How noisy is lexical decision? Front. Psychol. 3, 348

Murray, W.S. and Forster, K.I. (2004) Serial mechanisms in lexical access: the rank hypothesis. Psychol. Rev. 111, 721–756

Adelman, J.S. and Brown, G.D.A. (2008) Modeling lexical decision: the form of frequency and diversity effects. Psychol. Rev. 115, 214–229

Balota, D.A. et al. (2008) Beyond mean response latency: response time distributional analyses of semantic priming. J. Mem. Lang. 59, 495–523

Balota, D.A. and Spieler, D.H. (1999) Word frequency, repetition, and lexicality effects in word recognition tasks: beyond measures of central tendency. J. Exp. Psychol. Gen. 128, 32–55

Andrews, S. and Heathcote, A. (2001) Distinguishing common and task-specific processes in word identification: a matter of some moment? J. Exp. Psychol. Learn. Mem. Cogn. 27, 514–544

Wagenmakers, E-J. et al. (2008) A diffusion model account of criterion shifts in the lexical decision task. J. Mem. Lang. 58, 140–159

Gomez, P. et al. (2013) A diffusion model account of masked versus unmasked priming: are they qualitatively different? J. Exp. Psychol. Hum. Percept. Perform. http://dx.doi.org/10.1037/a0032333

Yap, M.J. et al. (2011) Individual differences in visual word recognition: insights from the English lexicon project. J. Exp. Psychol. Hum. Percept. Perform. 38, 53–79

Ratcliff, R. et al. (2004) A diffusion model analysis of the effects of aging in the lexical-decision task. Psychol. Aging 19, 278–289

Dufau, S. et al. (2012) How to say “na” to a nonword: a leaky competing accumulator model of lexical decision. J. Exp. Psychol. Learn. Mem. Cogn. 38, 1117–1128

Gomez, P. et al. (2008) The overlap model: a model of letter position coding. Psychol. Rev. 115, 577–600

Norris, D. and Kinoshita, S. (2008) Perception as evidence accumulation and Bayesian inference: insights from masked priming. J. Exp. Psychol. Gen. 137, 434–455

Norris, D. et al. (2010) A stimulus sampling theory of letter identity and order. J. Mem. Lang. 62, 254–271

Ratcliff, R. and McKoon, G. (2008) The diffusion decision model: theory and data for two-choice decision tasks. Neural Comput. 20, 873–922

Rayner, K. et al. (2006) Reading words with jumbled letters: there is a cost. Psychol. Sci. 17, 192–193

Velan, H. and Frost, R. (2007) Cambridge University versus Hebrew University: the impact of letter transposition on reading English and Hebrew. Psychon. Bull. Rev. 14, 913–918

Perea, M. and Lupker, S.J. (2003) Does jugde activate COURT? Transposed-letter similarity effects in masked associative priming. Mem. Cognit. 31, 829–841

Kinoshita, S. and Norris, D. (2009) Transposed-letter priming of prelexical orthographic representations. J. Exp. Psychol. Learn. Mem. Cogn. 35, 1–18

Perea, M. and Lupker, S. (2004) Can CANISO activate CASINO? Transposed-letter similarity effects with nonadjacent letter positions. J. Mem. Lang. 51, 231–246

Schoonbaert, S. and Grainger, J. (2004) Letter position coding in printed word perception: effects of repeated and transposed letters. Lang. Cogn. Process. 19, 333–367

Dehaene, S. et al. (2005) The neural code for written words: a proposal. Trends Cogn. Sci. 9, 335–341

Dehaene, S. (2009) Reading in the Brain. Viking

Grainger, J. et al. (2006) Letter position information and printed word perception: the relative-position priming constraint. J. Exp. Psychol. Hum. Percept. Perform. 32, 865–884

Davis, C. (2010) SOLAR versus SERIAL revisited. Eur. J. Cogn. Psychol. 22, 695–724

Chung, S. and Legge, G. (2009) Precision of position signals for letters. Vision Res. 49, 1948–1960

Peressotti, F. and Grainger, J. (1999) The role of letter identity and letter position in orthographic priming. Percept. Psychophys. 61, 691–706

Van Assche, E. and Grainger, J. (2006) A study of relative-position priming with superset primes. J. Exp. Psychol. Learn. Mem. Cogn. 32, 399–415

Welvaert, M. et al. (2008) Graded effects of number of inserted letters in superset priming. Exp. Psychol. 55, 54–63

Kinoshita, S. and Norris, D. (2013) Letter order is not coded by open bigrams. J. Mem. Lang. 69, 135–150

Roberts, S. and Pashler, H. (2000) How persuasive is a good fit? A comment on theory testing. Psychol. Rev. 107, 358–367

Lewandowsky, S. and Farrell, S. (2010) Computational Modeling in Cognition: Principles and Practice. Sage

Pitt, M.A. et al. (2008) Measuring model flexibility with parameter space partitioning: an introduction and application example. Cogn. Sci. 32, 1285–1303

Pitt, M.A. et al. (2006) Global model analysis by parameter space partitioning. Psychol. Rev. 113, 57–83

Jacobs, A.M. and Grainger, J. (1994) Models of visual word recognition: sampling the state of the art. J. Exp. Psychol. Hum. Percept. Perform. 20, 1311–1334

Roelofs, A. (2005) From Popper to Lakatos: a case for cumulative computational modeling. In Twenty-first Century Psycholinguistics: Four Cornerstones (Cutler, A., ed.), pp. 315–330, Erlbaum

Forster, K.I. (1994) Computational modeling and elementary process analysis in visual word recognition. J. Exp. Psychol. Hum. Percept. Perform. 20, 1292–1310

Norris, D. (2005) How do computational models help us develop better theories? In Twenty-first Century Psycholinguistics: Four Cornerstones (Cutler, A., ed.), pp. 331–346, Erlbaum

Griffiths, T.L. et al. (2010) Probabilistic models of cognition: exploring representations and inductive biases. Trends Cogn. Sci. 14, 357–364

Griffiths, T.L. et al. (2012) How the Bayesians got their beliefs (and what those beliefs actually are): comment on Bowers and Davis (2012). Psychol. Bull. 138, 415–422

Bex, G.E.P. and Draper, N.R. (1987) Empirical Model Building and Response Surfaces. John Wiley & Sons