



# Modeling the length effect for words in lexical decision: The role of visual attention

Emilie Ginestet, Thierry Phénix, Julien Diard, Sylviane Valdois

## ► To cite this version:

Emilie Ginestet, Thierry Phénix, Julien Diard, Sylviane Valdois. Modeling the length effect for words in lexical decision: The role of visual attention. Vision Research, 2019, 10.1016/j.visres.2019.03.003 . hal-02097508

**HAL Id: hal-02097508**

**<https://hal.science/hal-02097508>**

Submitted on 12 Apr 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Modeling the length effect for words in lexical decision: The role of visual attention

Emilie Ginestet<sup>1,2</sup>, Thierry Phénix<sup>1</sup>, Julien Diard<sup>1</sup> and Sylviane Valdois<sup>1</sup>

<sup>1</sup>*Univ. Grenoble Alpes, CNRS, LPNC, 38000, Grenoble, France*

<sup>2</sup>*corresponding author: emilie.ginestet@univ-grenoble-alpes.fr*

---

## Abstract

The word length effect in Lexical Decision (LD) has been studied in many behavioral experiments but no computational models has yet simulated this effect. We use a new Bayesian model of visual word recognition, the BRAID model, that simulates expert readers performance. BRAID integrates an attentional component modeled by a Gaussian probability distribution, a mechanism of lateral interference between adjacent letters and an acuity gradient, but no phonological component. We explored the role of visual attention on the word length effect using 1,200 French words from 4 to 11 letters. A series of five simulations was carried out to assess (a) the impact of a single attentional focus versus multiple shifts of attention on the word length effect and (b) how this effect is modulated by variations in the distribution of attention. Results show that the model successfully simulates the word length effect reported for humans in the French Lexicon Project when allowing multiple shifts of attention for longer words. The magnitude and direction of the effect can be modulated depending on the use of a uniform or narrow distribution of attention. The present study provides evidence that visual attention is critical for the recognition of single words and that a narrowing of the attention distribution might account for the exaggerated length effect reported in some reading disorders.

**Keywords:** Word length effect, Lexical decision, Bayesian modeling, Visual attention, Reading

---

## 1. Introduction

The effect of word length on visual word processing has been examined with a variety of techniques, including the lexical decision (LD) task in which the participant has to decide as quickly and as accurately as possible if the stimulus presented on the screen is a word or not. Initial research in this field yielded inconsistent results, the effect being in turn reported as inhibitory – longer words yield longer latencies (Balota et al., 2004; Hudson and Bergman, 1985; O'Regan and Jacobs, 1992) – or null (Acha and Perea, 2008; Frederiksen and Kroll, 1976; Richardson, 1976). In measuring LD reaction times (RT) for a large sample of words of different lengths, megastudies helped understand these apparent inconsistencies. Indeed, it appears that the effect of word length is not linear: reaction times are constant for words between 5 to 8 letters, but they increase with length for words longer than 8 letters (Ferrand et al., 2010, 2011, 2017; New et al., 2006).

Simulating the length effect in LD is challenging for all classes of word processing models, namely the dual-

route model, the triangle model or the multi-trace memory model. These models typically attribute the length effect to serial processing and can account for this effect in reading aloud but not in LD. According to dual route models (DRC: Coltheart et al., 2001; CDP, CDP+, Perry et al., 2007, 2010), an effect of length on naming latencies reflects serial processing within the non-lexical route, due to left-to-right grapheme parsing (Perry and Ziegler, 2002; Perry et al., 2007, 2010) or serial letter-sound mapping (Coltheart et al., 2001). Serial processing within the non-lexical route can straightforwardly account for the strong length effect typically reported in pseudo-word naming (Ferrand, 2000; Ferrand and New, 2003). The interaction between parallel processing by the lexical route and serial processing by the non-lexical route accounts for smaller length effects in word naming. Perry et al. (2007) showed that no length effect occurred when the non-lexical route of the CDP+ model was turned off, suggesting that serial processing is critical to explain length effects on words in naming. In a comparison of the length effect in English and German, Perry and Ziegler (2002) reported an effect of length on

naming latency in German but not in English for words from 3 to 6 letters. They simulated this differential effect within the DRC framework by changing the balance between lexical and non-lexical processing. Decreasing the strength of the lexical route while increasing that of the non-lexical route resulted in the observed length effect on German words, suggesting again that length effect would be inherently related to serial processing for phonological recoding. It follows that dual route models cannot account for the word length effect in LD, except by explaining it as following from serial phonological recoding, that is to say, assuming that the phonology of the word is generated in LD, as in naming, and that decision would be based on the phonological output.

The triangle model of reading (Seidenberg and McClelland, 1989; Plaut et al., 1996), that postulates parallel processing and does not include any serial mechanism, failed to simulate any length effect in word naming (Seidenberg and McClelland, 1989; Seidenberg and Plaut, 1998; Perry et al., 2007). In an attempt to accommodate the length effect on naming latency within this framework, Plaut (1999) implemented a serial processing mechanism that was orthographic and not phonological. The network initially fixated the first letter of the input string and tried to generate the appropriate phonological output of the whole word. When unable to generate the appropriate output based on this first fixation, the model had the ability to refixate the input string. Using the number of fixations as a proxy for naming latency, an overall length effect was obtained for words. This effect was null for 3-to-4 letter words that could be accurately identified within a single fixation, but inhibitory for 4-to-6 letter words that required more than one fixation. This study again emphasizes the critical role of serial processing for an account of length effects in word naming but provides no explanation on such effects in LD. Indeed, the number of refixations reflects the degree of difficulty that the network experiences in constructing word pronunciation without postulating any orthographic word recognition system that would be critical to simulate the LD task.

While the two previous classes of models assume involvement of phonological recoding in length effects, the multitrace memory (MTM) model of reading (Ans et al., 1998) postulates that such effects follow from visual attention processing at the orthographic level. The model postulates an attentional device, the visual attentional window, that delineates the amount of orthographic information under processing. Most familiar words are processed as a whole following single visual attention focusing, but shifts of focused attention are required to process unfamiliar letter strings. As a direct

consequence, simulations showed a strong length effect for pseudo-words in naming (Ans et al., 1998; Valdois et al., 2006). However, no length effect was simulated in LD for either words or pseudo-words, since decision was taken following parallel processing (i.e., single attention focusing) of the input letter-string (Valdois et al., 2006). A length effect in LD was simulated within this framework by assuming a reduced visual attentional window. The simulations carried out to account for impaired reading due to limited visual attention capacity showed significant length effect in both reading and LD and a stronger length effect for pseudo-words than for words (Juphard et al., 2004). Thus, the MTM model offers an account of the length effect as deriving from visual processing at the orthographic level, constrained by attention, but failed to simulate the length effect on words exhibited by typical readers in LD.

Overall, there is a relative consensus that the length effect, when observed, would reflect some kind of serial processing. What remains controversial is whether this serial mechanism relates to phonological decoding or to orthographic processing. Some behavioral data rather support an orthographic processing interpretation. Indeed, evidence that length effect can be seen in different word recognition tasks and that a larger effect is reported in tasks of progressive demasking, that more tap into visual factors than in lexical decision or naming tasks (Ferrand et al., 2011), suggests that this effect may reflect visual-orthographic encoding processes rather than orthography-to-phonology mapping. Furthermore, strong evidence for a visual-orthographic account comes from pathological data on brain-damaged patients and dyslexic children (Barton et al., 2014). The reports of an increased length effect in word reading following brain-damage in letter-by-letter readers (Arguin and Bub, 2005; Rayner and Johnson, 2005) or after surgical intervention at the level of visual-orthographic brain regions in patients with preserved oral language and phonological skills (Gaillard et al., 2006) support a visual-orthographic origin of the word length effect. In the same way, dyslexic children with a visual attention span disorder (i.e., a multiletter parallel processing deficit) show an abnormally large length effect in naming and lexical decision tasks, despite preserved phonological skills (Juphard et al., 2004; Valdois et al., 2003, 2011). These findings place important constraints on recognition models in suggesting that a visual attention mechanism may contribute to the word length effect.

The main contribution of the current paper is to use an original Bayesian model of visual word recognition, called BRAID (Phénix, 2018; Phénix et al., 2018) to study and simulate the word length effect in LD. The

BRAID model is a fully probabilistic model that includes a visual attention layer, an interference mechanism between adjacent letters and an acuity gradient. It was previously shown (Phénix, 2018) that it could account – as other models do – for classical effects in letter perception (e.g., word and pseudo-word superiority effects, context effects) and in word recognition and LD (e.g., frequency and neighborhood effects). We further showed that its visuo-attentional layer allowed to modulate some of these effects, accounting for more subtle effects such as letter spacing or cueing in letter perception. We also showed how its temporal nature allowed studying the dynamics of evidence accumulation about letters and words, allowing to reconcile seemingly contradictory effects of word neighborhood in LD (Phénix et al., 2018). Here, we use this model to simulate the length effect on LD latency for a large subset of words from the French Lexicon Project (FLP; Ferrand et al., 2010). We will specifically show the critical role of visual attention in the word length effect on LD latency.

The rest of this paper is structured in three main sections. In the first section, we describe the BRAID model and how LD is simulated by Bayesian inference. In the second section, we describe the stimuli used in the simulations and the method used to calibrate the simulated LD task. The third section reports five simulation experiments showing how the distribution of visual attention over the word letter string modulates word length effects in LD.

## 2. The BRAID model

The general structure of the BRAID model is illustrated in Figure 1. A full description of the model is provided elsewhere (Phénix, 2018), and beyond the scope of this paper. Instead, we propose a rapid summary of the salient points of the model, selected according to their relevance to the simulations carried out in the current study.

First, the overall architecture of the model is inspired from classical word recognition models (e.g., McClelland and Rumelhart, 1981), featuring three main representational levels. The first level, called, in our model, “the letter sensory submodel”, implements low-level visual processing of letter stimuli. These are noted, considering time instant  $T$ , with variables  $S_1^T$  to  $S_N^T$  (see Figure 1), with  $N$  the length of the input string. Processing at this level aims at recognizing letter identity and position (Grainger et al., 2016; Dehaene et al., 2005), which is also classical. Feature extraction is parallel and results in probability distribution over “internal” letter representations (variables  $I_1^T$  to  $I_N^T$  of Fig-

ure 1). The model implements several plausible components of low-level visual processing, such as an acuity gradient: information about letters decreases as distance from fixation position (variable  $G^T$  in Figure 1) increases. In probabilistic terms, this decrease of information is represented with increasing uncertainty in the corresponding probability distributions, of the form  $P(I_n^T | S_n^T)$ , which are identified from behavioral confusion matrices (Geyer, 1977). This acuity effect is symmetric around fixation (Whitney, 2001) and by default, we consider that gaze position  $G^T$  is located at word center. Concerning letter position identification, the model features a distributed position coding scheme (Gomez et al., 2008; Davis, 2010), such that information about a letter combines with neighboring letters. This mechanism is implemented by lateral interference between letters (represented with diagonal green edges in Figure 1).

The second submodel is the “letter perceptual submodel”. It implements how probability distributions over variables  $I_1^T$  to  $I_N^T$ , at each time step, feed information to be accumulated into perceptual variables  $P_1^T$  to  $P_N^T$  (see Figure 1); this creates an internal representation of the input letter string. The third level is the “lexical knowledge submodel”, implements knowledge about the spelling of 35,644 French words of the French Lexicon Project (Ferrand et al., 2010). For each word (variable  $W^T$  in Figure 1), its spelling is encoded as probability distributions over its letters (variables  $L_1^T$  to  $L_N^T$  in Figure 1). The probability to recognize a word is modulated by its frequency (prior probability distribution  $P(W^0)$ , not shown in Figure 1).

The letter perceptual and lexical knowledge submodels are linked by a layer of “coherence variables” (variables  $\lambda_{L_1}^T$  to  $\lambda_{L_N}^T$  in Figure 1), which allow, during word recognition or lexical decision, the comparison between the letter sequence currently perceived and those of known words. Coherence variables, here, can be interpreted as “Bayesian switches” (Gilet et al., 2011; Bessière et al., 2013). Depending on their state (open/closed or unspecified), the coherence variables allow or do not allow propagation of information between adjacent submodels. In that sense, they allow to connect or disconnect portions of the model. In the BRAID model, propagation of information through variables  $\lambda_{L_n}^T$  is bidirectional between the lexical knowledge and the letter perceptual submodels.

The BRAID features a fourth submodel, which is more original: the “visual attentional submodel”. It serves as an attentional filtering mechanism between the letter sensory submodel and the letter perceptual submodel. “Control variables” (variables  $C_1^T$  to  $C_N^T$  in Fig-

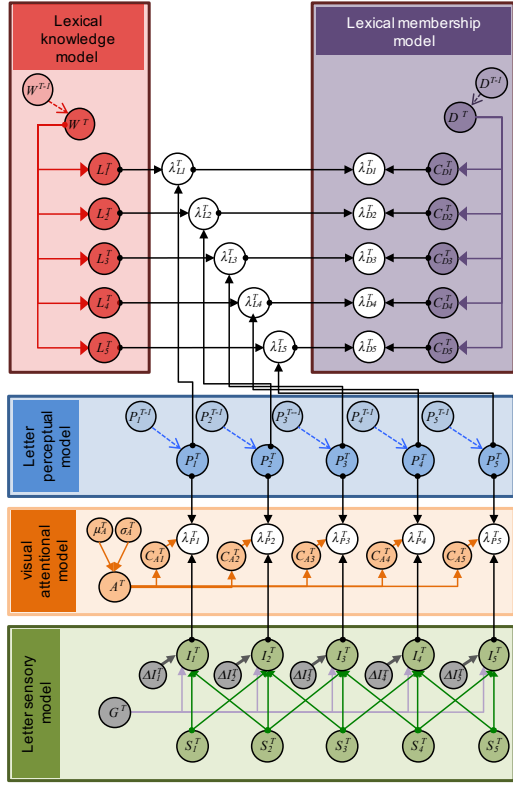


Figure 1: **Graphical representation of the structure of the BRAID model.** Submodels are represented as colored blocks, and group together variables of the model (nodes). The dependency structure of the model is represented by arrows. This dependency structure corresponds to a 5-letter stimulus (e.g., note the 5 spatial positions of variables  $S_1^T$  to  $S_5^T$ ); this structure also corresponds to time instant  $T$  (note that variables  $P_n^T$ ,  $W^T$  and  $D^T$ , on which dynamical models are defined, depend on their previous iterations,  $P_n^{T-1}$ ,  $W^{T-1}$  and  $D^{T-1}$ ). See text for details.

ure 1) pilot the states of coherence variables between the sensory and perceptual letter submodels (variables  $\lambda_{P_1^T}$  to  $\lambda_{P_N^T}$  in Figure 1): this allows explicitly controlling the transfer of information between these two submodels. We interpret this as an attentional model. We spatially constrain attention to be described by a probability distribution, so that sensory information cannot accumulate into perceptual variables in full, in all positions simultaneously: allocating attention to some positions is to the detriment of other positions. Mathematically, in the model, the distribution of attention over the letter string is Gaussian (variable  $A^T$  in Figure 1; the parameters of its distribution are its mean  $\mu_A^T$  and standard-deviation  $\sigma_A^T$ ). In this paper, we assume that the peak of attention is aligned with gaze position ( $\mu_A^T = G^T$ ). This model affects perceptual accumulation of evidence as acuity does: the further the letter from the attention

mean, the less attention it receives, the less information is transferred, and hence, the slower it is identified.

The fifth and final submodel is the “lexical membership submodel”. It implements knowledge about whether a sequence of letters (variables  $C_{D_1^T}$  to  $C_{D_N^T}$  in Figure 1) corresponds (variable  $D^T$  in Figure 1 is *true*) or not (variable  $D^T$  is *false*) to a known word. The knowledge encoded here can be interpreted as an “error model”: assuming the input string is a word, the perceived letters should match those of a known word in all positions; on the contrary, if the input is not a word, matching should fail in at least one position. This submodel is useful for simulating the lexical decision task.

## 2.1. Task simulation by Bayesian inference in BRAID

Variables appearing in Figure 1 form a state space with many dimensions. To mathematically define the BRAID model, the joint probability distribution over this state space is decomposed as a product of elementary probabilistic terms. Their definition and calibration is described in full elsewhere (Phénix, 2018). The model being defined, we then simulate tasks by computing a “question”, that is to say, a probability distribution of interest. This question is solved using Bayesian inference, that is to say, applying the rules of probabilistic calculus to the model. Here, we show the questions that allow simulating letter recognition, word recognition and lexical decision. The corresponding mathematical derivations cannot be provided here in full, due to lack of space. Instead, we describe how these derivations can be interpreted, in terms of simulated processes. The lexical decision task is the task of interest here, but since it involves the two previous ones, we describe them in their nesting order, for clarity purpose.

### 2.1.1. Letter identification

Consider first letter identification, that is, the process of sensory evidence accumulation, from a given stimulus, to perceptual letter identity. We distinguish variables and their values by using uppercase and lowercase notations, respectively. Furthermore, we use a shorthand for denoting all positions 1 to  $N$  and all time-steps 1 to  $T$  of any variable  $X$ :  $X_{1:N}^{1:T}$ . As an example, simulating that the model is given a sequence of letters as an input is setting variables  $S_{1:N}^{1:T}$  to  $s_{1:N}^{1:T}$ .

Thus, to simulate letter identification, we set stimulus  $s_{1:N}^{1:T}$ , gaze position  $g^{1:T}$  and attention parameters  $\mu_A^{1:T}$  and  $\sigma_A^{1:T}$ , we allow information propagation by setting variables  $\lambda_{P_n}^{1:T}$  to their “closed” state, and we compute the probability distribution, for a given time-step  $T$  and a given position  $n$ , over perceived letter  $P_n^T$ . Since

variables  $\lambda_L$  are left unspecified, information does not propagate to lexical submodels, and is constrained in the letter sensory, visual attentional and letter perceptual submodels (see Figure 1).

As a result, we consider here letter identification without lexical influence. We note that a variant in which variables  $\lambda_{L_{1:N}}^{1:T}$  are closed results in a bidirectional transfer of information, with both bottom-up perceptual influence of letters on words and top-down predictive influence of words on letters. This “letter identification with lexical influence” variant allows accounting for the classical word superiority effect (Ph  nix, 2018). We do not consider it further here; letter identification is therefore modeled by the question:

$$Q_{P_n^T} = P(P_n^T \mid s_{1:N}^{1:T} [\lambda_{P_n}^{1:T} = 1] \mu_A^{1:T} \sigma_A^{1:T} g^{1:T}). \quad (1)$$

$Q_{P_n^T}$  is a probability distribution over the perceived letter ( $P_n^T$ , at time  $T$  and position  $n$ ). It is a discrete distribution, since  $P_n^T$  is a variable with 27 possible values (one for each letter plus one for missing or unknown characters).

Computing  $Q_{P_n^T}$  involves two components, that are classical of inference in dynamical probabilistic models, such as Dynamic Bayesian Networks or Bayesian Filters (Bess  re et al., 2013). The first component is a dynamical prediction computation, whereas the second describes sensory evidence processing. In the dynamical term, the knowledge about letters at previous time step is spread to the next time step. This involves information decay such that, if stimuli were absent, the probability distribution over letters would decay towards its initial state. In the BRAID model, this is a uniform distribution, representing lack of information: all letters are considered equally likely at the perceptual level (i.e., lexical information is restricted to be expressed at the lexical knowledge submodel).

The second component describes sensory evidence processing at current time-step and its accumulation into the dynamic state variable. Here, information is extracted from stimulus  $s_{1:N}^T$ , in the letter sensory submodel, to accumulate into variable  $P_n^T$ . Details are not provided here, but this feature processing involves interference effects from adjacent letters, if any, and loss of performance, due to the combined effects of acuity and attention functions, when gaze and attention are not located on the considered letter position  $n$ .

However, the output of the letter sensory submodel is modulated by the visual attention submodel before reaching perceptual variable  $P_n^T$ . More precisely, attention allocation affects the balance between information decay and sensory evidence accumulation. We note  $\alpha_n$

the amount of attentional resources at position  $n$  (i.e., it is the probability value  $P([A^T = n] \mid \mu_A^T \sigma_A^T)$ ). When  $\alpha_n$  is high, sufficient attentional resources are allocated to position  $n$ , enough information from sensory processing accumulates into the perceptual variable so that temporal decay is counterbalanced and surpassed by sensory evidence. In this case, the probability distribution over variable  $P_n^T$  “acquires information” and gets more and more peaked at each time step. That peak, in the space of all possible letters, is always on the correct letter, provided enough attention (except for some pathological cases). In other words, letter identification, as simulated by the probability distribution  $Q_{P_n^T}$ , converges.

### 2.1.2. Word identification

Simulating word identification proceeds in a similar fashion as in isolated letter recognition above, except that information is allowed to propagate further into the model architecture, and more precisely to the lexical knowledge submodel, by setting  $[\lambda_{L_{1:N}}^{1:T} = 1]$  (see Figure 1). The probabilistic question is  $Q_W^T$ :

$$Q_W^T = P(W^T \mid e^{1:T} [\lambda_{L_{1:N}}^{1:T} = 1] \mu_A^{1:T} \sigma_A^{1:T}), \quad (2)$$

with  $e^t = s_{1:N}^t [\lambda_{P_{1:N}}^t = 1] g^t$ .

$Q_W^T$  is a probability distribution over words of a given lexicon; at any time-step  $T$ , it encodes information about which of these words is likely to correspond to the letter sequence perceived from the stimulus. As in letter identification, the resulting Bayesian inference (not detailed here, see Ph  nix, 2018) involves a classical structure, combining a dynamical system simulation with perceptual evidence accumulation. First, information in the probability distribution  $Q_W^T$  decays over time, so that, if the sensory stimuli were absent, it would converge back towards its resting and initial state. Here, this is the prior distribution over words  $P(W^0)$ , which encodes word frequency. Second, sensory evidence accumulation is based on the probabilistic comparison (through coherence variables  $\lambda_{L_1}^T$  to  $\lambda_{L_N}^T$ ) between letter sequences associated to words  $w$  in lexical knowledge and the perceived identities of letters, as computed by the letter recognition question  $Q_{P_n^T}$ . This comparison, in BRAID, is influenced by the similarity between the letter sequences of the stimulus and of words of the lexicon, so that similar (neighbor) words compete with each other for recognition.

This results in a dynamical process of word recognition that depends on letter recognition. In the first few time-steps,  $Q_{P_n^T}$  is still close to uniform, so that  $Q_W^T$  is, too. Sensory evidence accumulation then proceeds; after some time, and even though all letters might not

be perfectly identified yet (i.e., with high probability), the probability distributions over letters become diagnostic enough so that the input word is identified. The dynamics of this process are modulated, for instance, by the target word neighborhood density: recognition is faster when few perceptual evidence points toward a word with few or no competitors. Assuming that the input letter string corresponds to a known word, and with few pathological exceptions, word recognition as simulated by  $Q_W^T$  converges toward the correct word.

### 2.1.3. Lexical decision

The final task we consider is lexical decision, our task of interest for the experiments that follow. It is modeled by question  $Q_D^T$ :

$$Q_D^T = P(D^T \mid e^{1:T} [\lambda_{D_{1:N}}^{1:T} = 1] \mu_A^{1:T} \sigma_A^{1:T}) \quad (3)$$

$Q_D^T$  is a probability distribution, at each time-step, over the lexical membership variable  $D^T$ . It is a Boolean variable, that is to say, it is *true* when the stimulus is perceived to be a known word, and *false* otherwise. In question  $Q_D^T$ , the states of coherence variables allow information to propagate throughout the whole model, from the input letter-string to the lexical submodels (see Figure 1). Bayesian inference that derives from question  $Q_D^T$ , as before, involves a dynamical component (information decay towards a resting state) and perceptual evidence accumulation about lexical membership. However, what constitutes perceptual evidence, here, is less easily interpreted than in letter and word recognition. To explain, we consider, in turn, the two Boolean alternatives.

First, consider the hypothesis that the stimulus is a word (the  $D^T = \text{true}$  case). Perceptual evidence of the LD process is the process of word recognition. Evidence about lexical membership is the probability that coherence variables  $\lambda_{L_1}^T$  to  $\lambda_{L_N}^T$  detect a match between perceived letters and those of a known word. In other words, lexical decision proceeds as if the lexical knowledge submodel was “observing” the probability of a match between the letter perceptual and lexical knowledge submodels. When a known word can be reliably identified from the stimulus by word recognition (or when a set of neighbor words is activated enough), then coherence variables  $\lambda_{L_{1:N}}^T$  have high probability, indicating a match, so that the probability that  $D^T = \text{true}$  is high, indicating that the stimulus is recognized as being a known word.

Second, consider now the hypothesis that the stimulus is not a word (the  $D^T = \text{false}$  case). Here, the expected states of coherence variables  $\lambda_{L_{1:N}}^T$  indicates

that at least one letter of the stimulus should not match, when compared to known words. In other words, accumulating evidence from word recognition operates under the assumption that there would be one error in the stimulus, compared to known word forms. Technically, since it is unknown, all possible positions for the error have to be evaluated. For a given error position, word recognition  $Q_W^T$  is computed with the input letter string in all other positions, and alternative letters in the considered position. In other words, the stimulus is probably not a word if changing at least a letter of the stimulus is required to match it to a known word. Perceptual evidence accumulation in  $Q_D^T$  proceeds by pitting the two hypotheses against each other: lexical decision results from this competition. As in word recognition, which can recognize the input word even though its letter are not fully identified, lexical decision, in some cases, reaches high probability that the input is a known word before it is identified. The overall dynamics are further modulated by neighborhood density around the target word.

## 3. Method

Having described the general structure of the BRAID model and Bayesian inference involved in task simulation, we now describe the dataset that was used in the simulations and consider parameter calibration. The majority of the parameters of BRAID have been calibrated on independent data, and thus have default values that we used in previous simulations (Phénix, 2018). Here, we consider and calibrate the parameters that are specific to LD. For this purpose, we first explore the effect of these parameters on the simulated LD task to identify the parameter values that best fit the LD RTs reported for a large sample of French words in the FLP (Ferrand et al., 2010).

### 3.1. Material

The words used in these simulations were extracted from the FLP database (Ferrand et al., 2010). Given the great variability of the characteristics of the words of the FLP, we first chose to restrict the dataset to words from 4 to 11 letters. Then, we trimmed each list by removing words of very high frequency and those associated with high error rates. Finally, 150 words per length were randomly selected. Thus, our final word set consisted of 1,200 words from 4 to 11 letters, with a mean written frequency of 57.9 per million ranging from 20.3 to 104.2 per million and with a mean number of orthographic neighbors of 1.9 ranging from 0.4 to 5.1.

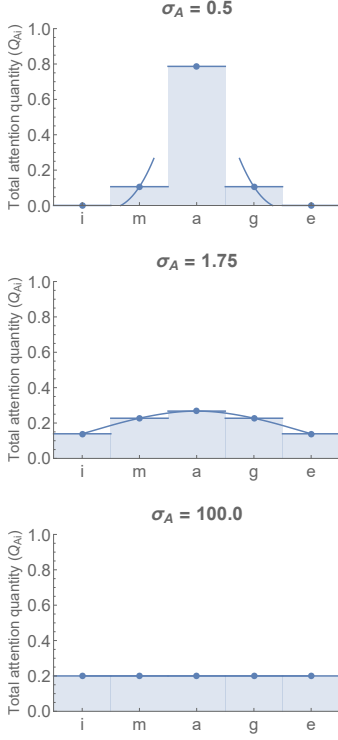


Figure 2: Illustration of attention distribution over the letter string of the word IMAGE for different values of the  $\sigma_A$  parameter: Left:  $\sigma_A = 0.5$ ; Middle:  $\sigma_A = 1.75$ ; Right:  $\sigma_A = 100.0$ .

An ANOVA on mean RTs for the selected word set showed significant main effect of length ( $F(7, 1173) = 2.342$ ,  $p = 0.022$ ) but no main effect of frequency ( $F < 1, ns$ ) and no Length-by-Frequency interaction ( $F < 1, ns$ ).

### 3.2. Calibration of parameters

We now present some of the parameters that have direct influence on simulating LD. First, we consider decision thresholds for the YES and NO answers,  $\tau_{YES}$  and  $\tau_{NO}$ . These parameters, with values between 0 and 1, are respectively linked to the probability  $P([D^T = true] | e^{1:T} [\lambda_{D_{1:N}}^{1:T} = 1] \mu_A^{1:T} \sigma_A^{1:T})$  that the stimulus is a word and to the probability  $P([D^T = false] | e^{1:T} [\lambda_{D_{1:N}}^{1:T} = 1] \mu_A^{1:T} \sigma_A^{1:T})$  that the stimulus is not a word. Thus,  $\tau_{YES}$  and  $\tau_{NO}$  set the probability to be reached by evidence accumulation to generate a decision. The relation between parameters  $\tau_{YES}$  and  $\tau_{NO}$ , as well as the values for the prior distribution over  $D^0$ , adapt the model to various types of non-words, according to their relation with real words (e.g., single-letter difference or full-consonant strings).

In the FLP experiment, whatever the word length, pseudowords were built in a single manner. Monosyllabic pseudowords were created by recombining onsets and rimes from the real words used. The same method was used to build polysyllabic pseudowords, but by recombining syllables instead of onsets and rimes. Despite controlling the same criteria as those used for the selection of real words, this method is questionable since it sometimes results in pseudowords that do not constitute plausible sequences in French. Thus, in our study, we only use the real words of the FLP experiment as stimuli. Therefore, we only consider simulations where words have to be identified as such (i.e., stimuli are always words), so that the parameter for deciding that a stimulus is not a word ( $\tau_{NO}$ ) should practically be irrelevant. To be precise, there could be fringe cases where a word would be incorrectly recognized as a non-word; such cases should be rare, especially for high values of  $\tau_{NO}$ .

Other parameters are easily set, thanks to their physical interpretations. For instance, we classically assume that the gaze position ( $g$ ) and the attentional focus ( $\mu_A$ ) coincide; thus, assuming a central fixation, we set  $\mu_A = g = \frac{N+1}{2}$ , with  $N$  the word length. The  $\sigma_A$  parameter characterizes the spread of attention, i.e., mathematically, the standard deviation of the Gaussian distribution of attention. Thus, the higher the value, the more spread out and uniform-like the distribution of attention. Since attention is modeled by a probability distribution, the sum of attention quantity  $Q_{Ai}$  over all letters is 1. Figure 2 illustrates various attention distributions for a given input. A reduced  $\sigma_A$  parameter value ( $\sigma_A = 0.5$ ) induces an attention distribution such that the central letter is efficiently processed, to the detriment of external letters. Conversely, a large  $\sigma_A$  parameter value ( $\sigma_A = 100.0$ ) spreads attention uniformly over all letters, such that they are equally processed but potentially insufficiently. Indeed, uniformly distributing a constant amount of attention will spread it too thinly for long words. Finally, an intermediate  $\sigma_A$  parameter value ( $\sigma_A = 1.75$ ) allows to distribute attention to favor some of the letters, but still provides enough processing resources to all the letters of a 5 letter word.

To specifically calibrate the LD related parameters  $\sigma_A$ ,  $\tau_{YES}$  and  $\tau_{NO}$ , we applied a grid search method, that is, we explored a set of regularly spaced points in the domain of possible parameter value combinations (including  $\tau_{NO}$  to verify that it is indeed irrelevant). We considered the grid search domain shown on Table 1.

The time unit of the BRAID model has been calibrated previously so that one iteration corresponds to

Table 1: Definition domains of  $\sigma_A$ ,  $\tau_{YES}$  and  $\tau_{NO}$  parameters for the calibration by grid search.

Parameters		Word length							
		4 L	5 L	6 L	7 L	8 L	9 L	10 L	11 L
$\sigma_A$	Min. value	0.5	0.5	0.5	0.5	1.0	1.0	1.0	1.0
	Max. value	3.25	3.25	3.25	3.25	4.0	4.0	4.5	4.5
	Step	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
$\tau_{YES}$	Min. value					0,6			
	Max. value					0,95			
	Step					0,05			
$\tau_{NO}$	Min. value					0,6			
	Max. value					0,95			
	Step					0,05			

around 1 ms. Thus, we set the maximum duration allocated for simulations to 2,000 iterations, which is enough time if we consider the mean RTs observed in the FLP (Ferrand et al., 2010).

Qualitative inspection of the results confirmed that the decision threshold  $\tau_{NO}$  has no influence on the performance of the model. In the following, we therefore set  $\tau_{NO}$  arbitrarily to 0.65, and only consider the values of couples  $(\tau_{YES}, \sigma_A)$  for further analyses.

To calibrate  $\sigma_A$  and  $\tau_{YES}$ , the model simulations were compared with the experimental observations of the FLP experiment. For this comparison, we used three measures: the Mean Square Error (MSE) between simulated and observed RTs for correct answers, the absolute difference between simulated and observed variances of RT distributions for correct answers, and, finally, the absolute difference between simulated and observed error rates. The MSE is calculated with:

$$MSE = \frac{1}{n} \sum_{i=1}^n (tc_i - tm_i)^2, \quad (4)$$

in which  $tc_i$  represents the observed RT (in ms) and  $tm_i$  the simulated RT (in number of iterations) for the  $i$ -th word, and  $n$  is the number of words correctly identified as words by the model. Results for each of these comparison measures, and for each explored  $(\tau_{YES}, \sigma_A)$  combination, are shown in Figure 3, aggregated for all word lengths (with qualitative inspection confirming that results behave consistently across word lengths).

As results suggest, each measure individually does not provide a unique point where model simulations and observations maximally correspond. Instead, each measure suggests a one-dimensional curve, along which combinations of  $\tau_{YES}$ ,  $\sigma_A$  provide good results. More precisely, increasing decision threshold  $\tau_{YES}$  can be compensated by also increasing  $\sigma_A$ , with a non-linear relationship in all three adequacy measures. This suggests that the model is robust, that is to say, a large num-

ber of parameter values allow the model to capture the experimental observations.

Combining the three adequacy measures was performed “geometrically” (Figure 3, bottom right), by considering the intersection of the three optimal curves of each measure. They intersect in a small region of the explored space, with  $\sigma_A$  between 1.6 and 1.8 and  $\tau_{YES}$  around 0.9.

In summary, simulating the LD task on 1,200 words from 4 to 11 letter allowed to calibrate three parameters ( $\sigma_A$ ,  $\tau_{YES}$  and  $\tau_{NO}$ ) of the BRAID model. Therefore, in the remainder of this study, we consider the following values:  $\tau_{NO} = 0.65$ ,  $\tau_{YES} = 0.90$  and  $\sigma_A = 1.75$ .

#### 4. Simulation 1: simulation of the word length effect

Above, in Section 3, we have aggregated simulation results over all word lengths, and shown that BRAID could account very well for the experimental observations, in terms of overall MSE, variance of response distributions, and accuracy. We now consider simulation results as a function of word length, to study the ability of the model to account for the word length effect. In this section, we study two simulations of the word length effect: in Simulation 1.A, as in parameter calibration, the model simulates LD with a single centered attentional focus, while in Simulation 1.B, the model performs several shifts of focused attention.

##### 4.1. Simulation 1.A

###### 4.1.1. Procedure

All parameter values used for this simulation are the same as those of parameter calibration (Section 3).

###### 4.1.2. Results

For analysis, the errors made by the model were removed, which, for the chosen parameters, represent 5

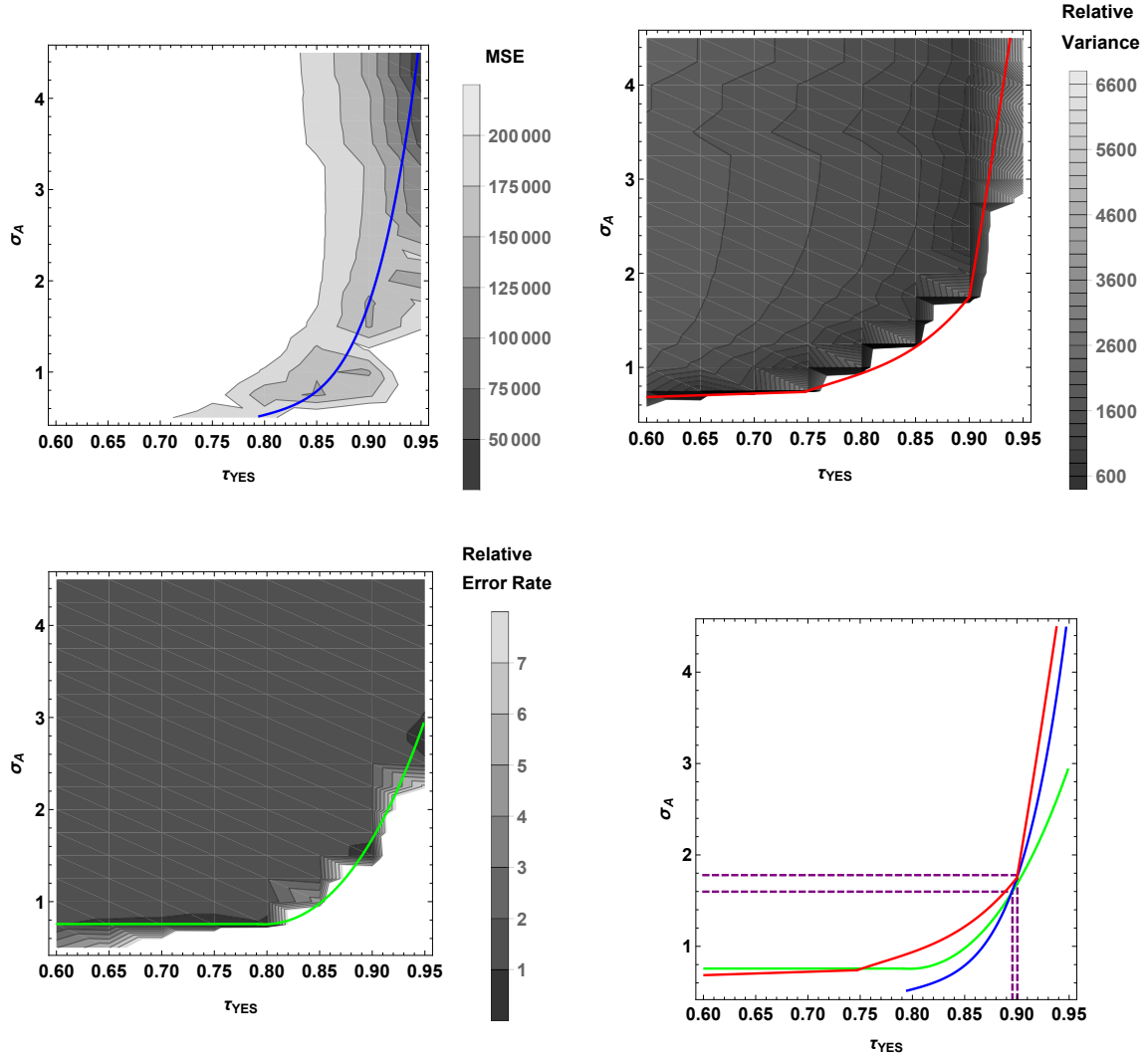


Figure 3: Results of the grid search calibration of  $\tau_{YES}$  and  $\sigma_A$ . Top left: MSE between simulated and observed RTs. Top right: absolute difference between simulated and observed variance of RT distributions. Bottom left: absolute difference between simulated and observed error rate. All measures are averaged over all considered word lengths (4 to 11 letters). Color gradients indicate measure values: the darker the color, the smaller the difference between the observed and the simulated data. Colored curves highlight parameter space regions where measure is close to optimal. Bottom right: superposition of the three curves of optimal parameter combinations, defining a small region where all measures are close to optimal (delineated by dashed lines).

words out of 1,200 (i.e., 0.42% of errors). Results are presented in Figure 4.

Figure 4 (left) shows mean behavioral and simulated RTs. To compare them, we scaled the simulated RTs by adding a constant, on the 4-letter word condition, so as to align them with behavioral RTs for this condition. Using this method, the simulation results for the other word lengths are predictions. Note that we will use this same method to present simulation results for all the following simulations.

Figure 4 (left) shows that the word length effect is larger in the model than in the behavioral data. The model simulates a substantial word length effect of 44.6 iterations per letter on average, well above the 7.25 ms per letter found in the behavioral data. The linear regression computed between the simulated and behavioral data (Figure 4, right) yields a high correlation coefficient ( $R^2 = 0.895$ ), which might suggest that the model, overall, adequately reproduces the variation of RTs as a function of word length. However, the linear regression

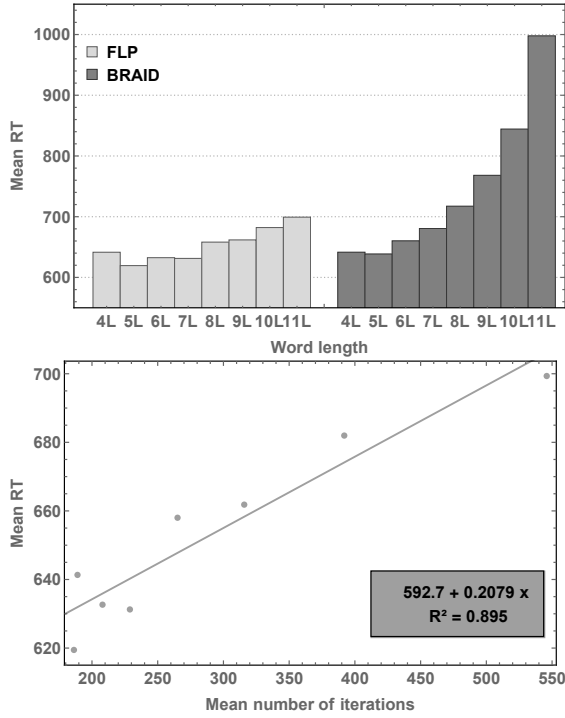


Figure 4: Results of Simulation 1.A (simulation of LD with a unique centered attentional focus). Left: mean RTs reported in the FLP study (in ms; light gray) and simulated by the model (in number of iterations; dark gray), as a function of word length. Mean simulated RTs are scaled and adjusted by aligning them on the 4-letter word condition; other length conditions are thus model predictions. Right: linear regression between simulated and observed mean RTs ( $R^2 = 0.895$ ).

parameters indicate that the relation between behavioral data and model simulations is mostly supported by the additional constant (592.7), thus decreasing strongly the weight of the model (0.2079).

## 4.2. Simulation 1.B

### 4.2.1. Procedure

The results of Simulation 1.A suggest that the model is able to reproduce the word length effect, but its magnitude is larger than observed experimentally. In Simulation 1.A, letters of long words are processed in a parallel manner, using a unique, central gaze and attentional focus. Attention distribution, in this case, allocates a small amount of processing resources to outer letters, basing word recognition essentially on a small number of inner letters; this makes it inefficient. Therefore, in Simulation 1.B, we explored a variant in which we assume that, for words of 7 letters or more, the model performs several shifts of attention.

Thus, we defined  $\mu_{A1} = g_1$  and  $\mu_{A2} = g_2$  as being two positions of the attention focus used by the model

during the LD task. We adapted these values systematically for each word length (Table 2), following the assumption that a long word would be treated as two short words. For instance, we suppose that an 8-letter word is processed as two 4-letter words. Therefore, applying the same calculations as before for central focus (see Section 3.2) yields  $\mu_{A1} = 2.5$  and  $\mu_{A2} = 6.5$ .

We set the maximum duration of the attention focus at each position to three conditions: 50, 100 and 200 iterations. Past this focus duration, the model shifts focused attention from this position to the other one, alternating either until the decision threshold or the time limit of 2,000 iterations is reached. All other parameter values are identical to those used in Simulation 1.A.

### 4.2.2. Results

No errors were made by the model in Simulation 1.B and the results were qualitatively similar, whatever the number of iterations allocated to each attention focus (50, 100 or 200). We here present the simulated length effects for the condition with 100 iterations.

Figure 5 presents a comparison of simulation results with experimental data. As previously, presented RTs are adjusted. Compared with Simulation 1.A (Figure 4), the main finding of Simulation 1.B is that allowing the model to perform several shifts of focused attention decreases RTs for words from 7 to 11 letters. The word length effect simulated by the model is characterized by a slope of 8 iterations per letter, very close to the experimentally observed slope in the FLP data (7.25 ms per letter). Linear regression of Simulation 1.B is also very satisfying ( $R^2 = 0.868$ ). Finally, coefficients of the linear regression (Figure 5, right) show that the weight of the model is close to 1 (1.13), reducing the value of the additional constant (414.5). This time constant could reflect motor response time, measured in the FLP but not modeled in BRAID.

An additional simulation was carried out to explore whether a good fit of the data was only found when forcing multiple shifts of attention for words up to 7 letters. Results of the simulations showed that allowing multiple fixations for all word lengths resulted in a very similar length effect of 9 ms per letter well within the range of length effects reported in experimental studies for word lexical decision.

## 5. Simulation 2: Effect of attention distribution

In previous sections, we simulated the word length effect with either one (Simulation 1.A) or two (Simulation 1.B) shifts of attention, and with attention parameters

Table 2: Fixation positions used for Simulation 1.B.

	4L	5L	6L	7L	8L	9L	10L	11L
$\mu_{A1} = g_1$	2.5	3.0	3.5	2.5	2.5	3.0	3.0	3.5
$\mu_{A2} = g_2$	-	-	-	5.5	6.5	7.0	8.0	8.5

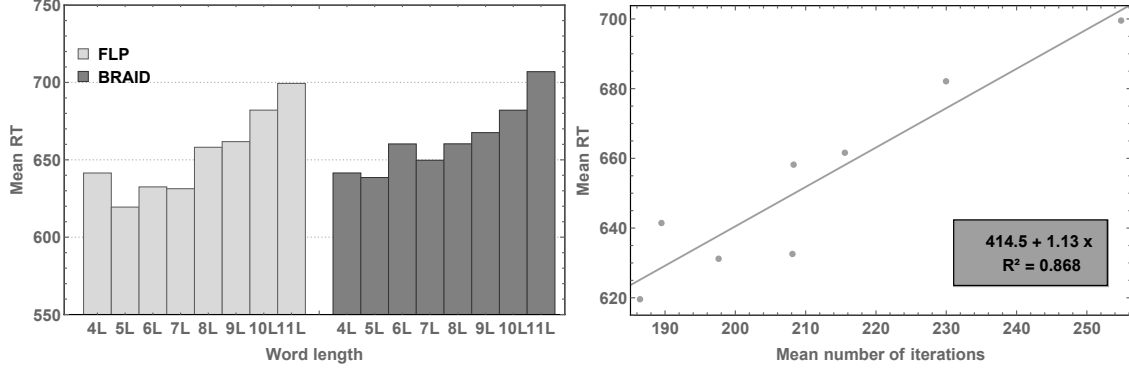


Figure 5: Results of Simulation 1.B (simulation of LD in the condition with two attention focuses up to 7 letter words and a focus duration of 100 iterations). Left: mean RTs reported in the FLP study (in ms; light gray) and simulated by the model (in number of iterations; dark gray) as a function of word length. Mean simulated RTs are scaled and adjusted by aligning them on the 4-letter word condition; other length conditions are thus model predictions. Right: linear regression between simulated and observed mean RTs ( $R^2 = 0.895$ ).

set to default values. We will now investigate the effect of attention dispersion, that is, the  $\sigma_A$  parameter controlling the standard deviation of attention distribution, on the simulated word length effect. For this, we perform three simulations (Simulation 2.A, 2.B and 2.C), using the same experimental material as previously.

In Simulations 2.A and 2.B, the model simulates a single central focus of attention, and allocates identical attentional resources to each letter, using uniform-like distributions of attention. In Simulation 2.A, the total amount of attention is stable across lengths while in Simulation 2.B, a similar amount of attention is arbitrarily allocated to each letter whatever word length (thus increasing total attention with word length). On the contrary, Simulation 2.C explores the other extreme case, in which attention is fully allocated to each letter, in turn, in a serial manner.

### 5.1. Procedure

Using our Gaussian model of attention distribution with  $\sigma_A = 100.0$  closely approximates a uniform distribution (see Figure 2, right). Recall that, in the model, the sum of the attentional quantity  $Q_{Ai}$  allocated to each letter is equal to 1. For this reason, the longer the word, the smaller  $Q_{Ai}$  for each letter. Thus, in Simulation 2.A, spreading attention uniformly gradually decreases the amount of attention allocated to each letter. Considering words from 4 to 11 letters, this yields  $0.09 \leq Q_{Ai} \leq 0.25$ .

It follows that the uniform distribution of a fixed amount of total attention interacts with word length, to the detriment of longer words. This manipulation of attention distribution thus predicts an effect of word length in Simulation 2.A. Simulation 2.B was performed to explore the impact of a uniform distribution of attention without a priori interacting with word length. For this purpose,  $Q_{Ai}$  was arbitrarily set to 0.5 whatever letter position and word length.

In Simulation 2.C,  $\sigma_A$  was set to 0.5 to simulate an extreme case of narrow attention distribution, in which attention is mostly allocated to a single letter at a time (see Figure 2, left). After 100 iterations on the first letter, the focus of attention shifts to the next letter and so on, in a left-to-right manner (cycling back to the first position after the last, if required), until either the LD threshold is reached or 2,000 iterations have passed. All the other parameters are identical to those used in previous simulations.

### 5.2. Results

No errors were made by the model in either Simulation 2.A or Simulation 2.B. Around five percent errors (64 words out of 1,200; 5.33% errors) that were removed for subsequent analyses occurred in Simulation 2.C. Figure 6 presents a comparison of simulation results with experimental data, in which, as previously, simulation results are adjusted by reference to the 4-letter word condition.

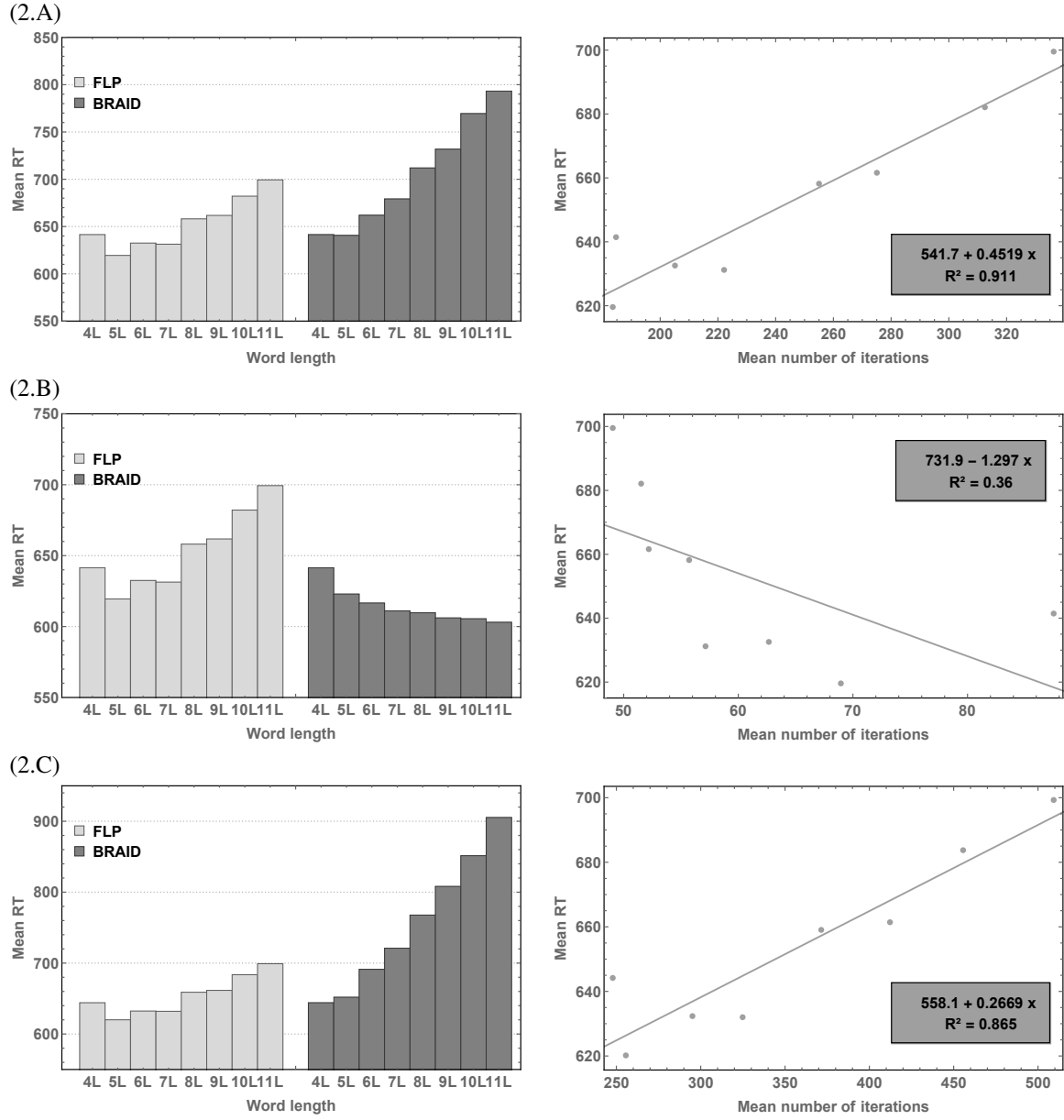


Figure 6: Results of the simulations of LD RTs for different conditions of attention distribution. Simulation 2A: uniform distribution of attention,  $Q_{Ai}$  varies from 0.25 to 0.09 depending on word length; Simulations 2.B: uniform attention distribution,  $Q_{Ai} = 0.5$  whatever word length; Simulation 2.C: narrow distribution of attention,  $\sigma_A = 0.5$ . Left: mean RTs obtained in the FLP study (in ms) and simulated by the model (in number of iterations) as a function of word length. Mean simulated RTs are scaled and adjusted by aligning them on the 4-letter word condition; other length conditions are model predictions. Right: linear regression between simulated (2.A, 2.B and 2.C) and observed mean RTs (respectively:  $R^2 = 0.911$ ,  $R^2 = 0.36$ ,  $R^2 = 0.865$ ).

As expected, Figure 6 (2.A, left) shows that the use of a uniform distribution of a stable amount of total attention across length resulted in a strong length effect on word LD RTs. An increase of 19 iterations per letter on average was obtained in the simulations against only 7.25 ms per letter in the behavioral data. The inhibitory

length effect obtained on word LD RTs in Simulation 2.A contrasts with the facilitatory effect observed in Simulation 2.B. Indeed, as shown on Figure 6 (2.B), longer words were processed faster when the amount of attention allocated to each letter was constant across positions. Figure 6 (2.C, left) presents results obtained for

Simulation 2.C. We observe that the magnitude of the simulated word length effect is larger than observed in the FLP study and this, starting from 6-letter words. The 32.6 iterations per letter found on average in the simulated data contrast with the 7.25 ms per letter reported in the behavioral data, showing that time of processing is highly and abnormally sensitive to word length.

The linear regressions presented on Figure 6 (2.A and 2.C, right) indicate that the relation between the behavioral and simulated data is almost linear (respectively  $R^2 = 0.911$  and  $R^2 = 0.865$ ). However, the parameters of the linear regressions, as in Simulation 1.A, show that these relations are mainly supported by the additional constant (respectively 541.7 and 558.1) decreasing strongly the weight of the model (respectively 0.4519 and 0.2669). Conversely, Figure 6 (2.B, right) shows that the relation between the behavioral and simulated data is not linear ( $R^2 = 0.36$ ).

## 6. Discussion

Our main purpose in this study was to provide new insights on the role of visual attention on the word length effect in lexical decision. The BRAID model was used to perform a series of simulations and compare their results with LD RTs for 1,200 words from 4 to 11 letters taken from the FLP (Ferrand et al., 2010). A first series of two simulations (Simulation 1.A and 1.B) was performed to simulate the behavioral word length effect in LD. In Simulation 1.A, length effect was simulated following a single central attention focus, which resulted in far stronger length effects than reported for humans in the FLP. In Simulation 1.B, a very good fit to the behavioral data was obtained using multiple shifts of attention during processing.

Another set of three simulations was then performed to manipulate the distribution of attention over the word letter-string and explore more in depth whether and how attention modulates the word length effect in LD. A uniform distribution of attention was used in Simulations 2.A and 2.B to explore processing in the absence of attention filtering. In Simulation 2.A, the total amount of attention remained constant across length so that each letter was allocated a lesser amount of attention as word length increased. In Simulation 2.B, the same amount of attention was allocated to each letter whatever word length. Results showed a stronger inhibitory word length effect than behaviorally reported, in Simulation 2.A, but a facilitatory effect in Simulation 2.B. Last, a Gaussian but narrow distribution of attention was used in Simulation 2.C that resulted in an exaggerated word length effect as compared to the behavioral data.

The current work shows that the BRAID model can successfully account for the length effect on word RTs in lexical decision, provided that multiple attention shifts across the letter-string are allowed for longer words. As previously claimed for length effects on naming latency (Plaut, 1999; Perry and Ziegler, 2002), this finding might suggest that length effects inherently relate to serial processing. Interestingly and rather counter-intuitively however, our simulations show that strong length effects can be generated with purely parallel processing (a single attention focus) and that, contrary to typical belief, additional shifts of attention lead to weaker rather than stronger length effects.

Indeed, comparing the results of the different simulations suggests that multiple shifts of attention fasten word processing, at least for longer words. Indeed, the most exaggerated word length effect on LD RTs was generated by BRAID following a single attention focus (Simulation 1.A). This effect was quite stronger than in conditions of multiple attention shifts, even as compared to the extreme case in which attention was allocated to each of the word letter successively (Simulation 2.C). Supportive evidence that length effect can follow from parallel processing comes from eye movement studies showing that the amount of time spent fixating a word increases with word length, even for single fixations (Rayner and Raney, 1996). Comparative studies of length effects in progressive masking and lexical decision provide further support to the present findings. While visual target display duration is not experimentally constrained in LD tasks, display time in progressive masking is short enough to prevent multiple fixations. Comparison of the two tasks showed that the word length effect was far stronger in progressive masking than in LD (Ferrand et al., 2011), suggesting larger length effects in conditions of parallel processing. Thus, a first contribution of the current work is to show that length effects cannot a priori be interpreted as direct evidence for serial processing.

A reliable simulation of the word length effect reported for humans on word LD RTs was here obtained assuming multiple attention shifts, thus multiple fixations, for words of 7 letters or more. Evidence that refixation probability and gaze duration increase with word length in conditions of text reading (Juhasz and Rayner, 2003; Pollatsek et al., 2008; Rayner and Raney, 1996; Rayner, 1998) or isolated word reading (Vitu et al., 1990), and that gaze duration relates with LD RTs (Schilling et al., 1998), makes plausible the prediction of BRAID of multiple refixations for longer words. However, multiple fixations have been reported on even shorter words (5-to-6 letter long) in isolated word read-

ing (Vitu et al., 1990), suggesting that multiple fixations might concern a wider range of word length than hypothesized in Simulation 1.B. Furthermore, the probability of a refixation is likely to gradually increase as a function of length. In contrast, we simulated a simple mechanism based on a length threshold: below it, refixations never happen; over it, refixations always occur. We also did not model the time delay induced by attention shifts. Modeling these mechanisms more precisely could help better account for the non-linearity of behavioral response times for words of lengths 5 to 7 (i.e., around our cutoff threshold for refixations in Simulation 1.B; see Figure 5). New experimental studies on eye movements during the lexical decision task are required to address these issues, as precise data on these mechanisms appear to be lacking. Overall however, the current work already shows that extending the possibility of multiple attention shifts/fixations to the entire word-length range (from 4 to 11 letters) only marginally affects the simulated word length effect on LD RTs. Accordingly, the capacity of BRAID to successfully account for the word length effect on LD RTs is quite robust provided that processing allows multiple attention shifts (thus, refixations). Overall, the current findings lead to conclude that refixating is beneficial to the reader – processing time is improved and reading is more fluent – which might explain why refixating is the rule in all languages and all semantic contexts, independently of the readers’ characteristics.

Another contribution of the current work is to show that length effects can be simulated in a word recognition model that does not include any phonological processing component. Previous computational accounts of the word length effect in isolated word naming attributed this effect to phonological recoding. For instance, in previous studies (Perry et al., 2007; Plaut, 1999), simulating the length effect was largely dependent on the capacity of the model to generate a plausible phonological output. As a direct consequence, such models cannot account for word length effects in lexical decision. Evidence in BRAID that mechanisms involved in visual word recognition can account for length effects independently of any phonological processing is well in line with evidence of word length effects in tasks such as progressive demasking, that primarily tap visual processing. This is also consistent with reports that the number of letters is a better predictor of LD latency than the number of phonemes (Balota et al., 2004). The fact that typical readers show exaggerated word length effects in conditions of visually degraded stimuli brings additional support for an early visual processing origin of word length effects in reading (Fiset et al., 2006).

This is not to say that length effects in isolated word naming exclusively derive from visual orthographic encoding skills. The current findings suggest that, in reading as in LD, word length effects might reflect early orthographic processing skills without excluding that additional effects due to phonological recoding may further affect word length effects in reading aloud.

Importantly, the main contribution of our work is to demonstrate the causal role of visual attention in LD word length effects. The whole set of simulations shows that variations in the distribution of visual attention over the input word has direct impact on word length effect patterns. The uniform distribution of a fixed amount of attention resources over the letter string led to an abnormally strong word length effect, which resulted from the combination of lower attentional resources and increased visual acuity effects for longer words. In contrast and despite the acuity gradient, a large and similar amount of attention allocated to each letter whatever word length led to faster LD latency for longer words, thus a reversed length effect as compared to the behavioral data. As a sidenote, these simulations also clearly show that neither visual acuity nor lateral interference (i.e., crowding) – which remain constant across simulations – are critical to account for word length effect. More importantly, the inability of BRAID to generate human-like word length effect through simulations based on uniform distributions of visual attention is further evidence that attention distribution is critical for word processing. The role of attention is certainly not to make the processing of multiple stimuli equal as illustrated in these simulations but rather to act as a filter that, as in Simulations 1.A and 1.B, enhances letter visibility in some portions of the word to the detriment of others (Carrasco, 2011). The capacity of BRAID to account for word length effect in LD primarily relies on a Gaussian distribution of attention, thus modulating attention across letters within the input word. Our account of visual attention as modulating letter processing within single words is strongly supported by experimental findings that show an impact of visual attention in single word processing (Besner et al., 2016; Lien et al., 2010) and lexical decision (McCann et al., 1992), that support early prelexical involvement of visual attention (Risko et al., 2010; Stolz and Stevanovski, 2004) and that assume modulation through visual attention of the rate of feature uptake (Stolz and Stevanovski, 2004; Carrasco, 2011). In BRAID, a fully defined visual attention device is for the first time implemented in a word recognition model, showing how attention modulates sensory processing and what is the impact of this modulation on word processing.

Last, another very important issue for a word recognition model is to account not only for typical but also for atypical reading. Although simulating patterns of acquired or developmental dyslexia was beyond the scope of the current paper, evidence suggests that a reduction of visual attention distribution over the letter string results in exaggerated word length effects (Barton et al., 2014; Duncan et al., 1999; Juphard et al., 2004; Valdois et al., 2003, 2011). Our Simulation 2.C, with larger length effects deriving from reduced attention dispersion, suggests that BRAID could be able to account for a variety of reading disorders that show exaggerated length effects in the context of visual attention deficits. Interestingly, providing an account of atypical word recognition skills was beyond the scope of most previous computational models of reading but the few that were concerned with pathology and tried to simulate acquired disorders did postulate an attentional device (Ans et al., 1998; Mozer and Behrmann, 1990). Recognizing the role of visual attention in single word recognition and reading and implementing a visual attention device in a computational model of reading like BRAID is critical to offer a plausible and integrative account of human reading skills.

## Acknowledgments

This work has been supported by a French Ministry of Research MESR grant for EG, and a Fondation de France research grant for TP.

## References

- Acha, J., Perea, M., 2008. The effects of length and transposed-letter similarity in lexical decision: Evidence with beginning, intermediate, and adult readers. *British Journal of Psychology* 99 (2), 245–264.
- Ans, B., Carbonnel, S., Valdois, S., 1998. A connectionist multiple-trace memory model for polysyllabic word reading. *Psychological Review* 105 (4), 678–723.
- Arguin, M., Bub, D., 2005. Parallel processing blocked by letter similarity in letter by letter dyslexia: A replication. *Cognitive Neuropsychology* 22 (5), 589–602.
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., Yap, M. J., 2004. Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General* 133 (2), 283.
- Barton, J. S., M, H. H., E, B. L., Charlotte, H., 2014. The word length effect in reading: A review. *Cognitive Neuropsychology* 31 (5–6), 378–412.
- Besner, D., Risko, E. F., Stolz, J. A., White, D., Reynolds, M., O'Malley, S., Robidoux, S., 2016. Varieties of attention: Their roles in visual word identification. *Current Directions in Psychological Science* 25 (3), 162–168.
- Bessière, P., Mazer, E., Ahuactzin, J. M., Mekhnacha, K., 2013. *Bayesian Programming*. CRC Press, Boca Raton, Florida.
- Carrasco, M., 2011. Visual attention: The past 25 years. *Vision Research* 51, 1484–1525.
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., Ziegler, J. C., 2001. DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review* 108 (1), 204–256.
- Davis, C. J., 2010. The spatial coding model of visual word identification. *Psychological Review* 117 (3), 713.
- Dehaene, S., Cohen, L., Sigman, M., Vinckier, F., 2005. The neural code for written words: a proposal. *Trends in Cognitive Sciences* 9 (7), 335–341.
- Duncan, J., Bundesen, C., Olson, A., Humphreys, G., Chavda, S., Chibuya, H., 1999. Systematic analysis of deficits in visual attention. *Journal of Experimental Psychology: General* 128 (4), 450–478.
- Ferrand, L., 2000. Reading aloud polysyllabic words and nonwords: The syllabic length effect reexamined. *Psychonomic Bulletin & Review* 7 (1), 142–148.
- Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., Augustinova, M., Pallier, C., 2011. Comparing word processing times in naming, lexical decision, and progressive demasking: Evidence from chronolex. *Frontiers in psychology* 2, 306.
- Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., Dufau, S., Mathôt, S., Grainger, J., 2017. MEGALEX: A megastudy of visual and auditory word recognition. *Behavior Research Methods*, 1–23.
- Ferrand, L., New, B., 2003. Syllabic length effects in visual word recognition and naming. *Acta Psychologica* 113 (2), 167–183.
- Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., Pallier, C., 2010. The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods* 42 (2), 488–496.
- Fiset, D., Gosselin, F., Blais, C., Arguin, M., 2006. Inducing letter-by-letter dyslexia in normal readers. *Journal of Cognitive Neuroscience* 18 (9), 1466–1476.
- Frederiksen, J. R., Kroll, J. F., 1976. Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance* 2 (3), 361.
- Gaillard, R., Naccache, L., Pinel, P., Clémenceau, S., Volle, E., Hasboun, D., Dupont, S., Baulac, M., Dehaene, S., Adam, C., et al., 2006. Direct intracranial, fMRI, and lesion evidence for the causal role of left inferotemporal cortex in reading. *Neuron* 50 (2), 191–204.
- Geyer, L., 1977. Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics* 22 (5), 487–490.
- Gilet, E., Diard, J., Bessière, P., 2011. Bayesian action-perception computational model: Interaction of production and recognition of cursive letters. *PLoS ONE* 6 (6), e20387.
- Gomez, P., Ratcliff, R., Perea, M., 2008. The Overlap model: a model of letter position coding. *Psychological Review* 115 (3), 577–601.
- Grainger, J., Dufau, S., Ziegler, J. C., 2016. A vision of reading. *Trends in Cognitive Sciences* 20 (3), 171–179.
- Hudson, P. T., Bergman, M. W., 1985. Lexical knowledge in word recognition: Word length and word frequency in naming and lexical decision tasks. *Journal of Memory and Language* 24 (1), 46–58.
- Juhász, B. J., Rayner, K., 2003. Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29 (6), 1312–1318.
- Juphard, A., Carbonnel, S., Valdois, S., 2004. Length effect in reading and lexical decision: Evidence from skilled readers and a developmental dyslexic participant. *Brain and Cognition* 55 (2), 332–340.
- Lien, M.-C., Ruthruff, E., Kouchi, S., Lachter, J., 2010. Event frequent and expected words are not identified without spatial attention. *Attention, Perception and Psychophysics* 72 (4), 973–988.
- McCann, R. S., Folk, C. L., Johnston, J. C., 1992. The role of spatial

- attention in visual word processing. *Journal of Experimental Psychology: Human Perception and Performance* 18 (4), 1015–1029.
- McClelland, J. L., Rumelhart, D. E., 1981. An interactive activation model of context effects in letter perception: Part I. an account of basic findings. *Psychological Review* 88 (5), 375–407.
- Mozer, M. C., Behrmann, M., 1990. On the interaction of selective attention and lexical knowledge: A connectionist account of neglect dyslexia. *Journal of Cognitive Neuroscience* 2 (2), 96–123.
- New, B., Ferrand, L., Pallier, C., Brysbaert, M., 2006. Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review* 13 (1), 45–52.
- O'Regan, J. K., Jacobs, A. M., 1992. Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology: Human Perception and Performance* 18 (1), 185–197.
- Perry, C., Ziegler, J. C., 2002. Cross-language computational investigation of the length effect in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance* 28 (4), 990–1001.
- Perry, C., Ziegler, J. C., Zorzi, M., 2007. Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. *Psychological Review* 114 (2), 273–315.
- Perry, C., Ziegler, J. C., Zorzi, M., 2010. Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology* 61 (2), 106–51.
- Phénix, T., 2018. Modélisation bayésienne algorithmique de la reconnaissance visuelle de mots et de l'attention visuelle. Ph.D. thesis, Univ. Grenoble Alpes.
- Phénix, T., Valdois, S., Diard, J., 2018. Reconciling opposite neighborhood frequency effects in lexical decision: Evidence from a novel probabilistic model of visual word recognition. In: Rogers, T., Rau, M., Zhu, X., Kalish, C. W. (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. Cognitive Science Society, Austin, TX, pp. 2238–2243.
- Plaut, D. C., 1999. A connectionist approach to word reading and acquired dyslexia: Extension to sequential processing. *Cognitive Science* 23 (4), 543–568.
- Plaut, D. C., McClelland, J. L., Seidenberg, M. S., Patterson, K., 1996. Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review* 103 (1), 56–115.
- Pollatsek, A., Juhasz, B. J., Reichle, E. D., Machacek, D., Rayner, K., 2008. Immediate and delayed effects of word frequency and word length on eye movements in reading: A reversed delayed effect of word length. *Journal of Experimental Psychology: Human Perception and Performance* 34 (3), 726–750.
- Rayner, K., 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124 (3), 372–422.
- Rayner, K., Johnson, R. L., 2005. Letter-by-letter acquired dyslexia is due to the serial encoding of letters. *Psychological Science* 16 (7), 530–534.
- Rayner, K., Raney, G. E., 1996. Eye movement control in reading and visual search: Effects of word frequency. *Psychonomic Bulletin & Review* 3 (2), 245–248.
- Richardson, J. T., 1976. The effects of stimulus attributes upon latency of word recognition. *British Journal of Psychology* 67 (3), 315–325.
- Risko, E. F., Stolz, J. A., Besner, D., 2010. Spatial attention modulates feature crosstalk in visual word processing. *Attention, Perception and Psychophysics* 72 (4), 989–998.
- Schilling, H. E., Rayner, K., Chumbley, J. I., 1998. Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition* 26 (6), 1270–1281.
- Seidenberg, M. S., McClelland, J. L., 1989. A distributed, developmental model of word recognition and naming. *Psychological Review* 96 (4), 523–568.
- Seidenberg, M. S., Plaut, D. C., 1998. Evaluating word-reading models at the item level: Matching the grain of theory and data. *Psychological Science* 9 (3), 234–237.
- Stolz, J. A., Stevanovski, B., 2004. Interactive activation in visual word recognition: Constraints imposed by the joint effect of spatial attention and semantics. *Journal of Experimental Psychology: Human Perception and Performance* 30, 1064–1076.
- Valdois, S., Bidet-Ildei, C., Lassus-Sangosse, D., Reilhac, C., Nguyen-Morel, M.-A., Guinet, E., Orliaguet, J.-P., 2011. A visual processing but no phonological disorder in a child with mixed dyslexia. *Cortex* 47 (10), 1197–1218.
- Valdois, S., Bosse, M.-L., Ans, B., Carbonnel, S., Zorman, M., David, D., Pellat, J., 2003. Phonological and visual processing deficits can dissociate in developmental dyslexia: Evidence from two case studies. *Reading and Writing: An Interdisciplinary Journal* 16, 541–572.
- Valdois, S., Carbonnel, S., Juphard, A., Baci, M., Ans, B., Peyrin, C., Segebarth, C., 2006. Polysyllabic pseudo-word processing in reading and lexical decision: converging evidence from behavioral data, connectionist simulations and functional MRI. *Brain Research* 1085 (1), 149–162.
- Vitu, F., O'Regan, J. K., Mittau, M., 1990. Optimal landing position in reading isolated words and continuous text. *Perception & Psychophysics* 47 (6), 583–600.
- Whitney, C., 2001. How the brain encodes the order of letters in a printed word: the SERIOL model and selective literature review. *Psychonomic Bulletin & Review* 8 (2), 221–43.