Title
A Fully Analytical Model of the Visual Lexical Decision Task

Author
Martin, Fermin Moscoso Del Prado

Peer reviewed

A Fully Analytical Model of the Visual Lexical Decision Task

Fermín Moscoso del Prado Martín (fermin.moscoso-del-prado@univ-provence.fr)
Laboratoire de Psychologie Cognitive (UMR 6146)
Aix-Marseille Université & Centre National de la Recherche Scientifique
Marseilles, France

Abstract

This study describes an analytical model of the visual lexical decision (VLD) task. The task is modeled as a problem of hypothesis testing given noisy evidence using a general Bayesian framework, similar to several previously published models. The Bayesian formulation is then shown to reduce to a Geometric Brownian Motion with a drift and an infinitesimal variance (a Drift-Diffusion Model). In turn, this reduction enables the use of direct analytical techniques – instead of simulations – to understand the different factors that influence response latencies. We demonstrate the power of this technique by analyzing the individual response latencies to a realistic size vocabulary covering virtually the full English lexicon. The model achieves an accurate prediction of the response latencies and error scores to thousands of individual words in relation to previously published VLD data. Crucially, this approach enables a direct explanation of several known non-linear effects, directly addressing their underlying mathematical explanation in a level of detail that is not attainable using traditional simulation-based approaches.

Keywords: Visual Lexical Decision; Bayesian; Analytical; Brownian Motion; English; Distributed Representations

A Bayesian Model of Lexical Decision

In line with some current models of the VLD task (e.g., Adelman & Brown, 2008; Norris, 2006; Ratcliff, Gómez, & McKoon, 2004; Wagenmakers, Steyvers, Raaijmakers, Shiffrin, van Rijn, & Zeelenberg, 2004) we can view lexical decision as a problem of optimally taking a decision by accummulating evidence coming from a noisy input. In VLD the input corresponds to the visual information provided by the eyes as time passes. As more samples of the input are accumulated the evidence supporting a 'yes' or a 'no' response grows until a certain level of certainty is attained. As noted by Adelman and Brown (2008), Norris (2006), or Wagenmakers et al. (2004), from a Bayesian perspective, the problem of deciding whether a certain visual input corresponds to a word or not can be characterized as a general problem of hypothesis choice. We can view word versus non-word decision as two hypotheses from which participants have to choose one, and assume that the participants respond as soon as they have gathered a certain level of evidence in favor of either.

Odds Ratios

The Odds Ratio (OR) between two hypotheses is the ratio of their posterior probabilities given the available information. In our case the two hypotheses being word (W) and non-word (NW), the decision on presentation of a certain visual input (I) could be made using the OR:

$$B = \frac{P(W|I,\mathcal{H})}{P(NW|I,\mathcal{H})}, \qquad (1)$$

where the additional conditionings on ($\mathcal{H}$) represent the set of modeling assumptions and previous knowledge under which we provide the estimates of the posteriors.

The input I corresponds to a sequence of samples $I = \{\mathbf{x}_1,\ldots,\mathbf{x}_T\}$ from a noisy distribution. Therefore we can view the OR in (1) as a function of time, whose value changes with each new sample that is received. In this way, W would be chosen over NW at the first time T when the value of the OR between them exceeds some threshold value $\Theta_W > 1$:

$$B(T) = \frac{P(W|\mathbf{x}_1,\ldots,\mathbf{x}_T,\mathcal{H})}{P(NW|\mathbf{x}_1,\ldots,\mathbf{x}_T,\mathcal{H})} \geq \Theta_W. \qquad (2)$$

This amounts to ensuring that the probability of W given the stimulus (and our assumptions) is at least Θ_W times greater than the probability of NW given the stimulus. Symmetrically, we can use another threshold $0 < \Theta_{NW} < \Theta_W$ that would enable us to choose to respond that the input corresponds to a non-word, whenever $B(T) \leq \Theta_{NW}$. These threshold parameters could take very different values, depending on how much easier we would expect one hypothesis to be recognized over another. However, for simplicity we will work under the assumptions that the thresholds are symmetrical ($\Theta_W = \Theta$, $\Theta_{NW} = \frac{1}{\Theta}$).

The calculation of the OR's is simplified when one works in a logarithmic scale, the Log Odds Ratio (LOR). If we define $\theta = \log\Theta$, then the condition to choose a word over a non-word stated in (2) is fully equivalent to:

$$b(T) = \log\frac{P(W|\mathbf{x}_1,\ldots,\mathbf{x}_T,\mathcal{H})}{P(NW|\mathbf{x}_1,\ldots,\mathbf{x}_T,\mathcal{H})} \geq \theta. \qquad (3)$$

The time T at which the condition in (2) is first satisfied is the same at which (3) becomes true.

By applying Bayes' Theorem on both terms of this ratio, we obtain the ratio of the likelihoods times the ratio of the prior probabilities. However, in a typical lexical decision experiment, the prior probability of observing a word or a pseudo-word are balanced and would normally cancel out. Thus we are left with the ratio between the likelihoods of the input sample $\{\mathbf{x}_1,\ldots,\mathbf{x}_t\}$ under a word or non-word hypothesis:

$$b(t) = \log\frac{P(\mathbf{x}_1,\ldots,\mathbf{x}_t|W,\mathcal{H})}{P(\mathbf{x}_1,\ldots,\mathbf{x}_t|NW,\mathcal{H})}. \qquad (4)$$

As in most models of this task, we assume that the samples from the visual input are independent of each other. Thus, we can expand this log-likelihood as the sum of the log-likelihoods of the independent samples ($\mathbf{x}_t$) from the input:

$$b(t) = \sum_{k=1}^{t} \log P(\mathbf{x}_k|W,\mathcal{H}) - \sum_{k=1}^{t} \log P(\mathbf{x}_k|NW,\mathcal{H}). \qquad (5)$$

This assumption of temporal independence of the samples could suffer if samples came from different eye fixations. However, on the lack of information of the specific fixations we can safely assume that their average can be described by a single distribution of independent samples.

Likelihood for Words

In an optimal decision process, it is not the likelihood of a particular word that drives the decision, but rather the combined likelihoods of all possible words, weighted by their individual prior probabilities:

$$P(I|W,\mathcal{H}) = \sum_{i=1}^{N_w} P(I|W_i,W,\mathcal{H})P(W_i|W,\mathcal{H}). \quad (6)$$

where N_w is the number of words in the lexicon. The first component of each of the terms of this sum, the prior for a particular word (W_i), is its overall probability of occurrence in the experiment. On the lack of additional contextual constraints, we can estimate it as being proportional to its relative frequency of occurrence in a linguistic corpus:

$$P(W_i|W,\mathcal{H}) \simeq \frac{F(W_i)}{\sum_{j=1}^{N_w} F(W_j)}. \quad (7)$$

In turn, the combined likelihoods given a word W_i for a sequence of t independent input samples $\mathbf{x}_1,\ldots,\mathbf{x}_t$ sampled from a multidimensional Gaussian with diagonal covariance ($\mathcal{N}(\mu_i,\sigma^2)$) is the product of the individual likelihoods for each of the $\{\mathbf{x}_1\ldots\mathbf{x}_t\}$:

$$P(\mathbf{x}_1,\ldots,\mathbf{x}_T|W_i,\mathcal{H}) = \prod_{k=1}^{T} P(\mathbf{x}_k|W_i,\mathcal{H}) \quad (8)$$

Each of the individual likelihoods in (8) is a Gaussian centered on μ_i and with a variance σ^2:

$$P(\mathbf{x}_k|W_i,\mathcal{H}) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\|\mathbf{x}_k-\mu_i\|^2}{2\sigma^2}}. \quad (9)$$

For simplicity, in line with previous models, we will assume that the sampling variance (σ^2), is uniform across experimental stimuli. This is a simplification of the actual process where different types of stimuli could give rise to different noisy variances (consider for instance the effect on the visual input to varying word lengths).

At this stage, we can already notice two factors that will affect the LOR. On the one hand, from (7) we can infer that the contribution that each word's contribution to the LOR will be proportional to its frequency of occurrence in a corpus. On the other hand, the likelihood in (9) decreases exponentially with the distance between each sampled input and the centroid of the representation of the target word. As the input itself will follow a normal distribution centered on this particular centroid, due to the product in (8) we can be certain that with time the contribution to the LOR will be mostly driven by the likelihood of the target word, with a minor contribution of relatively high frequency orthographic neighbors. Therefore this will give rise to a facilitatory effect of word frequency, as most of the value of the numerator to the LOR will come from the target word itself, and this contribution is frequency-weighted. In addition, at least at the initial stages of processing, the LOR will also receive a facilitatory contribution from the orthographic neighborhood, which will increase with the number, relative proximity, and frequency of existing neighbors.

Likelihood for pseudo-words

The problem is to decide whether the input has been sampled from a Gaussian centered in the mean of one of the existing words, or whether it is more likely to have been sampled from another Gaussian distribution with a different unknown mean, corresponding to a non-word.

A simple way to represent this is that the probability of a pseudo-word corresponds to the sum of the probabilities of possible non-words located at all points in the representational space, weighted by our prior expectation of finding a pseudo-word at that point in space:

$$P(\mathbf{x}_k|NW,\mathcal{H}) = \int_{-\infty}^{\infty} p(\mathbf{x}_k|\mathbf{m},NW,\mathcal{H})p(\mathbf{m}|NW,\mathcal{H})d\mathbf{m}, \quad (10)$$

where $\mathbf{m}$ are the possible locations in the representational where the pseudo-word could be located.[1]

The two components in this integral require assumptions on the corresponding distributions. As was done for the words, we can assume that the likelihood of observing a particular sample from the input $\mathbf{x}$ given that the presented stimulus was a non-word with orthographic representation $\mathbf{m}$ follows a multidimensional Gaussian centered on the representation of the non-word, and a diagonal covariance matrix with determiner σ^2:

$$p(\mathbf{x}_k|\mathbf{m},NW,\mathcal{H}) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\|\mathbf{x}_k-\mathbf{m}\|^2}{2\sigma^2}} \quad (11)$$

The non-words in a typical visual lexical decision experiment are constructed to be similar to the existing words, therefore their distribution in the representation space should be similar to that of the words themselves. On the lack of additional knowledge on the shape of distribution of words in the representational space, by the Maximum Entropy Principle we can assume it is a multidimensional Gaussian. We can estimate the frequency weighted mean of the representations of all words in our lexicon (μ_w) and the corresponding variance (σ_w^2). Therefore, the prior for the location of the possible pseudo-word means is:

$$p(\mathbf{m}|NW,\mathcal{H}) = \frac{1}{\sqrt{2\pi\sigma_w^2}}e^{-\frac{\|\mathbf{m}-\mu_w\|^2}{2\sigma_w^2}} \quad (12)$$

---

[1]This integral includes all locations in the representational space, also those of the words themselves. Note however, that in relation to the whole space, the combined cumulative probability of the points corresponding to words is zero, and thus negligible.

The integral in (10) can be reduced to a convolution between two Gaussians, thus it is itself also a Gaussian distribution with mean μ_w and variance $\sigma_w^2 + \sigma^2$:

$$P(I|NW, \mathcal{H}) = \frac{1}{\sqrt{2\pi(\sigma_w^2 + \sigma^2)}} e^{-\frac{\|\mathbf{x} - \mu_w\|^2}{2(\sigma_w^2 + \sigma^2)}}. \qquad (13)$$

In practice, the variance of the words over the whole lexicon is much greater than the variance of the input, $\sigma_w^2 \gg \sigma^2$. This implies that the expression in (13) will be very close to the original prior on the pseudo-word mean expressed in (12). The likelihood of the input given a pseudo-word changes little with time, and depends only on the centrality of the input in the lexical representational space. In general, the closer the input is to the center of the representational space (μ_w), the greater that the likelihood for the pseudo-word will be, and the smaller the value of the LOR. The center of the representational space is the area where a greater number of existing words can be expected: a denser orthographic neighborhood. Thus, when the input is coming from a dense orthographic neighborhood, the recognition of a word will be more difficult. It will become slower. This implies that the orthographic neighborhood size effect (and neighborhood frequency as well, as μ_w was computed as by weighting the contribution of words by their frequency) will have an inhibitory component in visual lexical decision, at least when the pseudo-words used in the experiment were designed to be similar to the words. This inhibitory component will in turn be combined with the facilitatory neighborhood density component that was discussed above.

Response Latencies

Distribution of the LOR in time

As discussed above, the time taken to decide whether the input corresponds to a word or to a non-word will be the first time T when the LOR in (5) reaches a value greater then θ ('yes' response) or lower than $-\theta$ ('no' response). Therefore the average RT to a particular word should correspond to the expected time that it takes the LOR to cross the θ barrier, without having previously crossed the $-\theta$ barrier (which would have already led to a 'no' response).

Closed-form expressions for the mean and variance of the ORs between the word and non-word hypotheses at each point in time can be obtained by integration. Unfortunately, the equations that one obtains in this way are computationally intractable. Using a motivated approximation, we can obtain an estimate of the reaction times. The value of the likelihood for the word hypothesis is mostly driven by the likelihood of the particular word that was presented, with a rather small contribution of the other words in the lexicon, which will be stronger at the earlier stages of processing. We can thus consider individually the contribution to the likelihood of the target word (W_i), and summarize in a single term the contribution of all other words (which is in any case minor). Our estimation of the likelihood of a pseudo-word used

a Gaussian distribution to simulate the probability of finding a pseudo-word in the different parts of the lexicon. This Gaussian was chosen to replicate as closely as possible the distribution of words in the lexicon. Therefore it is reasonable to employ this same distribution to *approximate* the contribution of the *other* words in the lexicon. The contribution of those others should be relatively smaller than that of the pseudo-words because the pseudo-words could be located anywhere in the space, while the existing words only occupy a few of those infinite possible locations. We therefore account for the contribution of the other words using a parameter $0 \le \alpha \le 1$. The approximated OR becomes:

$$\hat{B}(t) = P(W_i|\mathcal{H}) \frac{P(\mathbf{x}_1, \dots, \mathbf{x}_t | W_i, \mathcal{H})}{P(\mathbf{x}_1, \dots, \mathbf{x}_t | NW, \mathcal{H})} + (1 - P(W_i|\mathcal{H})) \alpha. \qquad (14)$$

Note that the term corresponding to the non-word likelihood cancels out.

With this simplification, the condition in (2) can be restated in terms or the LOR between one particular word (the one presented) and the non-words:

$$b'(t) = \log P(W_i|\mathcal{H}) \frac{P(\mathbf{x}_1, \dots, \mathbf{x}_t | W_i, \mathcal{H})}{P(\mathbf{x}_1, \dots, \mathbf{x}_t | NW, \mathcal{H})}, \qquad (15)$$

$$b'(t) \ge \log[\Theta - (1 - P(W_i|\mathcal{H}))\alpha]. \qquad (16)$$

Note that we have now a case of an asymmetrical threshold. The 'yes' decisions will be take with the threshold $\theta'_W = \log[\Theta - (1 - P(W_i|\mathcal{H}))\alpha]$, and the 'no' will use a threshold $\theta'_{NW} = \log\left[\frac{1}{\Theta} - (1 - P(W_i|\mathcal{H}))\alpha\right]$. Furthermore, the equations give us an additional constraint on the possible values of the α parameter. In order to be useful, the OR in (16) should never have a negative value. Therefore $\alpha \le \frac{1}{\Theta}$. To ensure the maximum possible contribution of the other words to the decision process, it is safest to assume that the parameter gets its maximum possible value, $\alpha = \frac{1}{\Theta}$.

The LOR $b'(t)$ between two possible hypotheses follows a normal distribution (c.f., Kass & Raftery, 1995). If we now integrate to calculate the expectation of $b'(t)$, we find that the expected value of the LOR at any time t is a linear expression of t:

$$E(b'(t)) = K + \upsilon \cdot t, \qquad (17)$$

where with K and υ have the values:

$$K = \log P(W_i|\mathcal{H}), \qquad (18)$$

$$\upsilon = \frac{1}{2} \log \frac{\sigma_w^2 + \sigma^2}{\sigma^2} + \frac{1}{2}\left(\frac{\sigma^2}{\sigma_w^2} - 1\right) + \frac{d_i^2}{2(\sigma_w^2 + \sigma^2)}, \qquad (19)$$

and d_i is the Euclidean distance between the prototypical representation of the presented word (μ_i) and the center of the representational space (μ_w):

$$d_i = \|\mu_i - \mu_w\| \qquad (20)$$

Furthermore, integration also reveals that the variance of the LOR also follows a (very similar) linear function of time.

Reaction Times

As described above, the value of the LOR between two words follows a Gaussian distribution, whose mean and standard deviation are a linear function of time. Instead of considering a discrete intake of samples from the distribution at a fixed rate – as was done by Adelman & Brown (2008) or Norris (2006) – we can consider the equivalent limiting process in which samples from the input are collected *continuously* in time. This limit continuous process is a *Brownian Motion* with a starting value (18), a drift (19), and an infinitesimal variance (the expected variance of the LOR). If we momentarily ignore the times when the negative threshold is reached first (i.e., the errors), the distribution of the times for such a process to reach a particular positive threshold value corresponds to the distribution of *first-passage times* of the Brownian motion through a fixed positive barrier. This distribution is known in its closed form: First-passage times of a Brownian motion follow an *Inverse Gaussian* distribution (IG). If we have a Brownian motion with a starting value K and a positive drift υ, the expected first passage time of the process through a positive level θ'_W is given by:

$$E(T_W) = \frac{\theta' - K}{\upsilon}. \qquad (21)$$

This expresses the intuitive notion that the average time to reach a preset level of certainty starting at time zero from an offset equal to our prior expectations, is equal to the difference between the desired level to be attained and the initial offset, divided by the average accumulation of evidence per unit of time.

The IG distribution describes the first-passage times through the positive threshold. However, it does not consider whether at that moment the negative threshold has already been crossed, in which case an error would have happened and the time would not affect our distribution. What we are interested in is in the distribution of the time taken to cross the positive threshold, provided that the positive one is crossed before the negative. This is expressed by the conditional probability function $p(T_W|\text{CorrectResponse}, W_i, \mathcal{H})$. We can use Bayes' theorem to calculate this distribution, and then integrate to find the corrected distribution of latencies (see Dixit, 1993 for a detailed discussion of these issues). Fortunately, in our particular case, introducing this correction did not produce significantly different results than those produced by just applying (21), so for simplicity we do not consider it in the remaining discussion.

Model Implementation

Orthographic Representations

In order to obtain estimates of the reaction times using our method, we need a distributed representation of the orthographic forms of all English words. For this purpose, we used the Accumulation of Expectations (AoE) technique (Moscoso del Prado Martín, 2003; Moscoso del Prado Martín, Ernestus & Baayen, 2004;

Moscoso del Prado Martín, Schreuder & Baayen, 2004). This technique enables us to automatically build distributed vectors representing all orthographic forms in a given language. These vectors have a fixed dimensionality, and do not require alignment of the words at their beginnings, or endings (40 dimensions per vector, for all words). The AoE vectors have successfully been used in large-scale connectionist models of the processing of Dutch and English words. We used the English vectors of Moscoso del Prado and colleagues to estimate the distribution of words in the English lexicon. For this, we employed the vectors corresponding to all English words appearing in the CELEX database (Baayen, Piepenbrock & Gulikers, 1995) with a frequency greater than one. In order to adapt them to the needs of our model, several modifications were done on these vectors. First, to ensure that the similarity space is defined by the Euclidean distance (the vectors were originally developed for use with angular measures), we normalized them to modulus one. Second, as reported by Moscoso del Prado Martín, (2003), these vectors tend to represent longer words in central areas of the representational space, which can lead to reversed word length effect. We overcome this problem by linearly scaling the vectors by their word length. Finally, in order to ensure the required diagonality of the covariance matrix, the vectors were rotated using a Principal Component Analysis[2]. We used these vectors to compute the frequency-weighted mean ($\mu_w \simeq \mathbf{0}$) and the determiner of the corresponding covariance matrix ($\sigma_w^2 = 14.74$) to use with the equations defined above.

Dataset and Model Fit

We investigated how accurately would our model predict VLD RT's of a previously published dataset. For simplicity, we chose a subset of the data described by Balota, Cortese and Pilotti (1999) for which a highly detailed analysis of the RTs was provided by Baayen et al., (2006). This subset contained the average young participants' VLD responses to $2,088$ monosyllabic mono-morphemic English words. Using the formulation from the previous section, we computed the predicted average VLD RT using a geometric Brownian Motion with an absorbing barrier. The drifts, infinitesimal variances, and biases were computed directly. The values of the two free parameters of the model the threshold Θ and the variance of the input error (σ) were set in different ways. On the one hand, the value of Θ was set using a Gauss-Newton nonlinear least-squares regression from the theoretical to the actual RTs. The value of σ was chosen to be small ($\sigma = .1$) relative to the variance of the words in the lexicon ($\sigma_w = 3.84$). We chose this value because it is the point were the parameter seemed to reach an asymptote in the prediction of reaction times (in general, the smaller this parameter, the better the prediction)[3].

---

[2]These transformed orthographic vectors can be obtained by contacting the author.

[3]We excluded σ from the non-linear regression because including it led to non-convergence of the regression algorithm, as the performance keeps improving infinitesimally as its value decreases. In
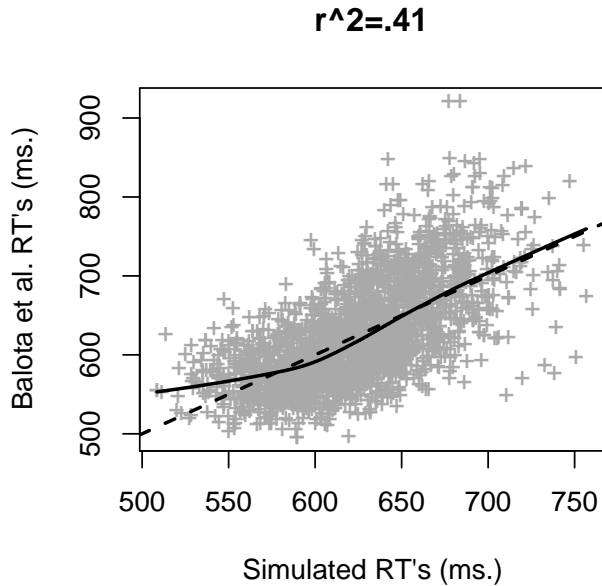
r^2=.41

Figure 1: Comparison of the theoretical RT's predicted by the model (horizontal axis), with the average VLD RT's of young participants to the same items in the Balota et al. (1999) study (vertical axis). The dashed line plots the identity relation. The solid line is a non-parametric regression between both measures.

Results and Discussion

Figure 1 shows the relationship between the predicted average response latencies using our model (horizontal axis), and the actual average response latencies of subjects performing VLD, taken from the Balota et al. (1999) dataset. The first thing to notice is that, despite the very large number of items, the model achieves a fairly good prediction of each individual latency, with an overall explained variance of over 40%. As illustrated by the overlap between the non-parametric regression (solid line) and the identity line (dashed line), the relation between both measures is remarkably linear. This indicates that the model captures the detailed distribution of response latencies, with enough detail to make item-level predictions.

In addition, one can observe a slight nonlinearity for the fastest reaction times. Although relatively small, this deviation from linearity is very robust, and will show up consistently in re-sampling analyses of this dataset. Interestingly, if we were to swap the axes, and perform the non-parametric regression in the opposite direction (i.e., predicting the model responses from the experimental RTs), we would find the same deviation from linearity in the low range. Counterintuitively, this deviation goes in the *same* direction (above

addition, a general linear scaling factor was added to speed up the convergence. Finally and additional intercept fixed to 447ms. was added to the model based on a separate theoretical study on the (relatively) constant portion of the VLD RT distributions.
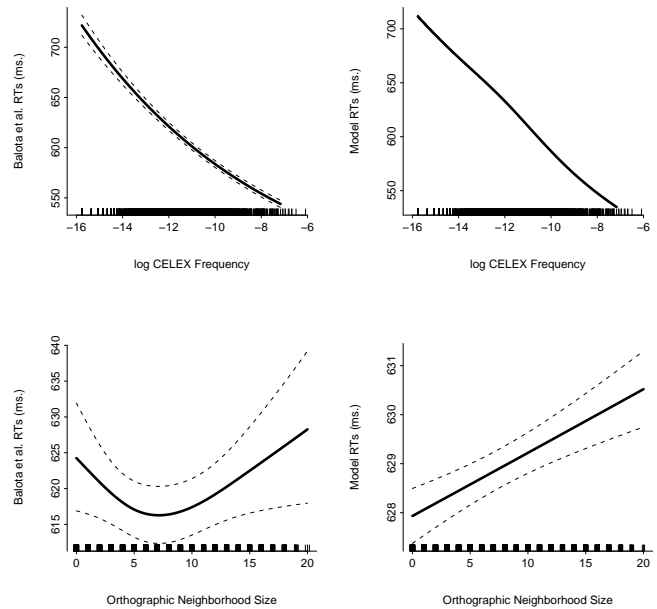

Figure 2: Comparison of the effects of word frequency (top panels) and orthographic neighborhood size (bottom panels) on the Balota et al. (1999) RT's (left panels), and on the RTs predicted by the model (right panels). The effects were estimated using a least-squares regression analysis on the log RT's including non-linear terms considered using restricted cubic splines. The rugs at the bottom of each panel illustrate the densities of the counts.

the diagonal), irrespective of the direction in which we performed the regression. In fact, rather than pointing to a deviation between the model's predictions and the participants' responses, this is an intrinsic property of the reaction times distribution. If we were to plot the RTs of individual subjects against the others, we would again find the same nonlinearity. This is no more than a form of regression towards the mean, which in very left-skewed distributions (as that of the RTs), is more marked in the lower than in the upper range (Baayen, Moscoso del Prado Martín, Schreuder, & Wurm, 2003).

We now turn to examine the role of frequency and neighborhood size of the experimental response latencies and model predictions. We performed a least-squares regression on both the experimental and the theoretical RT's, including log frequency and the orthographic neighborhood size variable, and including the possibility of a restricted cubic splines for the effects for which the non-linear term reached significance. Figure 2 summarizes the effects that we observed in these regressions.

First, with respect to frequency (top two panels), we observe that the both in the human RT's (left panel) and model responses (right panel) it played a clearly linear effect. In-

deed it is not surprising that frequency has a linear effect on the model predictions. From (21) we can see that log frequency will have a direct linear effect. Both Adelman and Brown (2008) and Norris (2006) defended the crucial role of log frequency on lexical decision latencies. In the case of Adelman and Brown they could infer that the evidence for log frequency was superior to the evidence for other counts like rank frequency (Forster, 1976). In Norris' study, the author had to conclude that both these counts could be equally good predictors for a Bayesian model of lexical decision. Notice that our analytical approach enables to conclude for certain that – at least on this type of models – it is log frequency rather than rank frequency that will drive the decision.

Finally, we come to the issue of orthographic neighborhood size effects. On the reaction times, we observed a non-linear contribution of this variable. Notice that the model, on the other hand, does not show this non-linearity, but rather shows a constant inhibitory effect of the variable on the predicted latencies. As we discussed when introducing the likelihoods for words and non-words, both of these will contribute in opposite directions to the neighborhood size effect, with the facilitation provided by the words being stronger at the earlier stages of processing. This is indeed the pattern that we observe on the human reaction times. In order to approximate the word likelihood, we made the contribution of the 'other' words constant in time (i.e., the α parameter), thus eliminating the non-linearity on the reaction times, which passes to be dominated just by the inhibition provided by the non-word likelihood. In turn, the direction of this effect is opposite to the effect reported by Norris (2006) on his model. As he assumed non-words to be uniformly distributed in the lexical space, there would be no reason in his model to expect that centrality and non-word density are correlated, thus eliminating the inhibitory contribution of the pseudo-words. Although Norris claims that the actual distribution of pseudo-words in the model matters little, in fact we can see that it is crucial for issues like neighborhood size. In fact, we can predict that this variable will also be affected by the distribution of the non-words in a real experiment.

In sum, we have introduced a fully analytical model of the VLD task (that can be run in seconds on a mid-range laptop). Although further work on this model is clearly necessary, as far as we are aware, this model outperforms any published model both in terms of coverage (around 40K words vocabulary), and item-level performance. Furthermore, by eschewing simulations, we can arrive at full-fledge analytical *explanations* of effects, rather than relying simply on goodness of fit statistics. The model described here has made use of a particular representational technique to represent the orthographic variation. Note however that, while the simulated results might depend on this particular representational scheme, the theoretical analysis holds for any representation of the visual form of words for which a vector-space can be defined, as long sampling is made based on the distance measure that defines the space.

References

Adelman, J. S., & Brown, G. D. A. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*, *115*, 214–227.

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*, 290–313.

Baayen, R. H., Moscoso del Prado Martín, F., Schreuder, R., L., & Wurm. (2003). When word frequencies may NOT regress towards the mean. In R. H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing.* Berlin: Mouton de Gruyter.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Balota, D., Cortese, M., & Pilotti, M. (1999). Item-level analyses of lexical decision performance: Results from a mega-study. *Abstracts of the 40th Annual Meeting of the Psychonomics Society, Psychonomic Society, Los Angeles*, 44.

Dixit, A. K. (1993). *The art of smooth pasting*. Chur, Switzerland: Harwood Academic Publishers.

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. C. T. Walker (Eds.), *New approaches to language mechanisms.* Amsterdam: North Holland.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Moscoso del Prado Martín, F. (2003). *Paradigmatic effects in morphological processing: Computational and cross-linguistic experimental studies*. Unpublished doctoral dissertation, Max Planck Institute for Psycholinguistics & University of Nijmegen.

Moscoso del Prado Martín, F., Ernestus, M., & Baayen, R. H. (2004). Do type and token effects reflect different mechanisms: Connectionist modelling of Dutch past-tense formation and final devoicing. *Brain and Language*, *90*, 287–298.

Moscoso del Prado Martín, F., Schreuder, R., & Baayen, R. (2004). Using the structure found in time: Building real-scale orthographic and phonetic representations by Accumulation of Expectations. In H. Bowman & C. Labiouse (Eds.), *Models of Cognition, Perception and Emotion. Proceedings of the VIII Neural Computation and Psychology Workshop.* Singapore: World Scientific.

Norris, D. (2006). The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*, 327–357.

Ratcliff, R., Gómez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 159–182.

Wagenmakers, E. J., Steyvers, M., Raaijmakers, J. G., Shiffrin, R. M., Rijn, H. van, & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology*, *48*, 332-367.