How trial-to-trial learning shapes mappings in the mental lexicon: Modelling Lexical Decision with Linear Discriminative Learning

Maria Heitmeier*, Yu-Ying Chuang and Harald Baayen
Quantitative Linguistics, University of Tübingen

arXiv:2207.00430v1 [cs.CL] 1 Jul 2022

Author's Note: This is a preprint. This paper has not been peer reviewed.

Priming and antipriming can be modelled with error-driven learning (Marsolek, 2008), by assuming that the learning of the prime influences processing of the target stimulus. This implies that participants are continuously learning in priming studies, and predicts that they are also learning in each trial of other psycholinguistic experiments. This study investigates whether trial-to-trial learning can be detected in lexical decision experiments. We used the Discriminative Lexicon Model (DLM; Baayen et al., 2019), a model of the mental lexicon with meaning representations from distributional semantics, which models incremental learning with the Widrow-Hoff rule. We used data from the British Lexicon Project (BLP; Keuleers et al., 2012) and simulated the lexical decision experiment with the DLM on a trial-by-trial basis for each subject individually. Then, reaction times for words and nonwords were predicted with Generalised Additive Models, using measures derived from the DLM simulations as predictors. Models were developed with the data of two subjects and tested on all other subjects. We extracted measures from two simulations for each subject (one with learning updates between trials and one without), and used them as input to two GAMs. Learning-based models showed better model fit than the non-learning ones for the majority of subjects. Our measures also provided insights into lexical processing and enabled us to explore individual differences with Linear Mixed Models. This demonstrates the potential of the DLM to model behavioural data and leads to the conclusion that trial-to-trial learning can indeed be detected in psycholinguistic experiments.

Keywords: trial-to-trial learning, linear discriminative learning, lexical decision, distributional semantics, mental lexicon

*maria.heitmeier@uni-tuebingen.de

1 Introduction

When going through our daily lives, we are constantly confronted with new information. What we see, hear and feel continuously updates our internal model of the world. This kind of continuous learning shapes how we perceive, process, learn and react to the world (e.g. Nassar et al., 2010; Bennett et al., 2015; O'Reilly et al., 2021; Diedrichsen et al., 2010; Ramscar et al., 2014, 2017; Ramscar, 2016; O'Reilly and Rohrlich, 2018). Learning does not only change our perception at a general level, but it also has immediate consequences for how we react to the world given what we have just perceived or experienced. Experimentally, this effect can be observed for example in repetition priming: after processing some information, when similar information is encountered again, it is processed more easily, which usually results in e.g. shorter reaction times compared to non-repeated information (see Roediger, 1993, for a review). Analogously, it has been found across many domains that the opposite is also true: if a repeated or similar stimulus is followed by a different outcome, processing is impaired, resulting in longer reaction times (an effect often referred to as "antipriming"; overview in Marsolek, 2008).

It has been found in recent work that priming effects can be modelled with a simple error-driven learning rule, called the Rescorla-Wagner learning rule (Rescorla and Wagner, 1972; Hoppe et al., 2022; Marsolek, 2008; Oppenheim et al., 2010; Baayen and Smolka, 2020). Error-driven learning, as modelled by the Rescorla-Wagner rule, assumes that when perceiving a input (often referred to as a *cue*), activations of outcomes are predicted. Then, the error between the actual outcome and its predicted activation are computed, and the mapping from cue to the observed outcome is strengthened accordingly. Mappings from the cue to all other outcomes that were activated but not observed are weakened. This mechanism accounts for priming: after perceiving a cue and an outcome, the mapping from cue to outcome is strengthened. When the cue is perceived again, the outcome is activated more strongly — resulting e.g. in shorter reaction times. At the same time, error-driven learning also provides an account of antipriming: mappings to activated outcomes which are not encountered with the cue are weakened, thus, reacting to a different outcome given the same cue becomes slower. The Rescorla-Wagner rule has recently been applied successfully to language learning and subsequently in many areas of psycholinguistics. In its simplest forms, it has been used to account for issues in (second) language acquisition (Ellis, 2006a,b; Ellis and Sagarra, 2010; Arnon and Ramscar, 2012; Hoppe et al., 2020) and ageing research (Ramscar et al., 2014, 2017), as well as semantic priming (Oppenheim et al., 2010), morphological processing (Baayen et al., 2011), learning of symbolic knowledge (Ramscar et al., 2010), the U-shaped learning of irregular English plurals in children (Ramscar and Yarlett, 2007; Ramscar et al., 2013) and early infant sound acquisition (Nixon and Tomaschek, 2021).

Modelling priming with the Rescorla-Wagner rule assumes that the learning of the prime influences processing of the target stimulus. This implies that participants are continuously learning in priming studies, and predicts that they are also learning in each trial of other psycholinguistic experiments. However, models examining lexical trial-to-trial effects in priming and other psycholinguistic studies so far have various shortcomings. The experimental stimuli in these simulations are typically carefully controlled for overlapping features between different cues. They have generally been designed for modelling specific purposes, and are often limited to a few items (e.g. Oppenheim et al., 2010; Lentz et al., 2021; Ramscar et al., 2013; Tomaschek et al., 2022). Models such as ACT-R (Anderson and Lebiere, 1998) model the learning and forgetting of stimuli also during experiments, but they treat words as units and model forgetting as a function of time (Van Rijn and Anderson, 2003), without taking into account interference caused by the learning of intervening stimuli, which is a crucial characteristic of the Rescorla-Wagner rule. Finally, Chang et al. (2006) modelled within-experiment learning using error-driven learning but focused on syntactic priming rather than lexical priming.

Within the current study, we therefore explore the effect of continuous learning with a model of the mental lexicon called the Discriminative Lexicon Model (DLM), and its learning mecha-

nism, Linear Discriminative Learning (LDL). The DLM posits simple mappings between representations of words' forms and their meanings (Baayen et al., 2018, 2019). The DLM has been successful both in modelling different morphological systems across a range of languages, such as Latin, English, German, Estonian, Korean and Maltese (Baayen et al., 2018, 2019; Chuang et al., 2020a, 2022a; Nieder et al., 2021), but at the same time also at modelling a range of behavioural data (Cassani et al., 2019; Heitmeier and Baayen, 2020; Chuang et al., 2020b; Shafaei-Bajestan et al., 2021; Heitmeier et al., 2021; Stein and Plag, 2021; Schmitz et al., 2021). It implements learning using an error-driven learning rule for continuous data (Widrow and Hoff, 1960) which is closely related to the later developed Rescorla-Wagner rule. Additionally, in contrast to previous models such as Naive Discriminative Learning (NDL; Baayen et al., 2011), it uses word embeddings to represent words' semantics. Word embeddings (aka semantic vectors) represent meanings in a distributed manner, building on the hypothesis that similar words occur in similar contexts (Harris, 1954). They are able to capture fine-grained meaning similarities between words and have been shown to predict numerous aspects of human processing in various studies (e.g. Baroni et al., 2014; Mandera et al., 2017; Westbury and Wurm, 2022; Westbury et al., 2014; Baayen et al., 2019).

The challenge addressed in this study is whether LDL is powerful enough to approximate continuous word learning. In contrast to previous work (Oppenheim et al., 2010; Lentz et al., 2021), we do not use a carefully controlled experiment designed to measure the effect of learning explicitly: instead, our hypothesis is that learning takes place during classical psycholinguistic experiments such as lexical decision, and that if this is indeed the case, the consequences of trial-to-trial learning can be modeled with LDL.

In a lexical decision task, participants have to decide whether a presented stimulus is an existing word in their language or not. Lexical decision is traditionally employed to probe the mental lexicon, measuring e.g. access speed. Being a traditional psycholinguistic experiment, megastudies of lexical decision are available, which have recorded lexical decision data by hundreds of participants for thousands of experimental stimuli for various languages, such as English, Dutch or Spanish (Balota et al., 2007; Keuleers et al., 2012; Brysbaert et al., 2016; Aguasvivas et al., 2018). In the present work, we use data from the British Lexicon Project (BLP; Keuleers et al., 2012), which encompasses lexical decision data from 78 participants for about 28,000 words and an equal number of nonwords. With datasets as big as these, even small effects of trial-to-trial learning should be detectable.

To develop our model and assess its ability to predict lexical decision reaction times, we proceed as follows: First, we explore how to develop the model and the statistical analyses with the data of subjects 1 and 2 (henceforth "training subjects"), which together cover all words and nonwords. We then regard the remaining 76 subjects as replication experiments against which we test the theory developed for subjects 1 and 2. This procedure ensures that a new theory is indeed predictive and does not overfit to the analysed data (see also Wilson and Collins, 2019).

In order to test our main hypothesis that during psycholinguistic experiments continuous learning can be detected, we proceed as follows: We use two instances of the DLM to predict participants' lexical decision reaction times: one with learning updates of the lexicon after each trial and one without. Then we test which of the models shows better fit to reaction times. If the model with incremental updates shows better model fit, we can conclude that continuous learning may indeed be taking place during the experiment.

Additionally to this main question, we also explore two further issues. First, we examine what the model tells us about processing in the mental lexicon in general. The representations and learning mechanisms that we are using in the present study have been found to be useful to predict behavioural data in previous work (e.g. Chuang et al., 2020b; Stein and Plag, 2021; Schmitz et al., 2021). We compare the measures extracted from the DLM with classical psycholinguistic predictors such as orthographic neighbourhood density. Our study therefore contributes to what Chuang and Baayen (2021) termed "external validation", i.e. measuring

the model's performance against behavioural data. Secondly, we explore individual differences. Previous work has shown that there are considerable individual differences in word recognition. For example, Kuperman and Van Dyke (2011) observed that in highly skilled readers, the frequency of the base word of morphologically complex words predicted longer reading latencies, whereas in low-skilled readers, it predicted shorter ones. Orthographic effects also differ across individuals. Milin et al. (2017a) conducted a serial reaction time experiment which they also simulated with NDL. They found that readers who speed up more across the experiment are less influenced by how much the target word is predicted by its orthographical cues than other subjects. Further studies confirm the influence of individual differences (e.g. Fischer-Baum et al., 2018; Perfetti et al., 2005), but note that connecting differences in morphological processing to individual psychological measures is difficult (Lõo et al., 2019). In the present work we focus on exploring individual differences in processing in the mental lexicon.

The paper is structured as follows: Section 2 gives an overview over previous computational models of lexical decision, and how the DLM relates to them. Section 3 introduces the DLM and Section 4 explains how lexical decision is modelled in the framework of the DLM. In Section 5 we give details on data preprocessing and the statistical models we employed to answer our main research questions. Section 6 reports our findings regarding insights into the mental lexicon which we can gain from the DLM, the effect of trial-to-trial learning as well as individual differences. Finally, Section 7 discusses our results.

2 Computational models of Lexical Decision

There exists a multitude of models of word recognition and lexical decision, beginning from so-called "box-and-arrow" models, which describe the processing of stimuli only verbally, all the way to full-fledged computational models. The latter set of models has the advantage that they need to specify each aspect of the model precisely and that they can predict behavioural data quantitatively, resulting in models which can be tested rigorously (e.g. Dell and Caramazza, 2008; McClelland, 2009; Bröker and Ramscar, 2020). This section gives a short overview of the most influential computational models which have been used to account for lexical decision, before contrasting them with the present approach.

Norris (2013) classifies computational models of reading and word recognition into different "styles" such as interactive activation (IA), mathematical-computational, and connectionist models. IA models are essentially networks with three different feature levels: letter features, letters, and words, implemented as nodes in the network. Nodes typically inhibit other nodes at their own level, and activate or inhibit nodes at higher levels. In order to recognise a word, first, relevant letter features are activated, which in turn activate letters which finally lead to activation of a word node fitting best to the activated letters (Rumelhart and McClelland, 1982). Models based on the original IA model usually took this basic architecture for granted and refined single aspects ("nested modelling", Jacobs and Grainger, 1994), such as the Spatial Coding Model (Davis, 2010), the Dual Route Cascaded Model (Coltheart et al., 2001) or the Multiple Read-Out model (Grainger and Jacobs, 1996). IA models are commonly initialised by assigning resting activation levels to the individual nodes. For word nodes these can be derived from word frequencies (McClelland and Rumelhart, 1989, Chapter 7). The original versions of the three models mentioned here did not include an account of learning, but learning mechanisms were developed for some of the later iterations of these models (e.g. Pritchard et al., 2016).

The second group, mathematical-computational models, are generally defined by mathematical functions rather than a network. The Diffusion Model (Ratcliff et al., 2004; Wagenmakers et al., 2008) is such a model. The model takes frequency and type of nonword as given, and uses these to let the response drift slowly either to a word or nonword response, the aim being to account for the distribution of reaction times in lexical decision. The model's parameters are usually either set by the modeller or estimated from existing data. The Bayesian Reader

(Norris, 2006) makes use of Bayes' formula to integrate the prior probability for various strings to be words (based on word frequency) with the incoming information on the target string to predict whether the string is a word or not.

A third style of models are so-called connectionist models. These models employ distributed representations rather than localist representations, and they usually make use of backpropagation of error (Rumelhart et al., 1986) to estimate the optimal connection weights. The use of distributed representations makes it possible to model fine-grained meaning similarities and differences. One example of an influential connectionist model is the triangle model (Seidenberg and McClelland, 1989; Harm and Seidenberg, 2004), which consists of orthography, phonology and semantic representations. The model can be trained, i.e. it "learns", and lexical decisions have been based on the error scores in this mapping (Seidenberg and McClelland, 1989). The model was later implemented as a recurrent neural network to enable the modelling of reaction times (Chang et al., 2013).

Lastly, a more recent style of modelling has emerged which Norris (2013) calls symbolic/localist models: Naive Discriminative Learning (NDL). NDL posits mappings between distributed representations of form (for different modalities) and meaning; instead of using backpropagation it makes use of the simplest form of error-driven learning, the Rescorla-Wagner rule (Rescorla and Wagner, 1972; Schultz, 1998; Marsolek, 2008; Trimmer et al., 2012; Ramscar et al., 2013; Hoppe et al., 2022), or the equilibrium equations of Danks (2003). The framework has been used to model both primed and unprimed lexical decision reaction times (Baayen et al., 2011; Milin et al., 2017b; Baayen and Smolka, 2020). Milin et al. (2017b) used an extension of the model where the localist meaning representation is understood as a pointer to a distributed meaning representation. Properties of this second embedding network were found to also be highly predictive for lexical decision times (Baayen et al., 2016).

In a pilot study, Chuang and Baayen (2021) used the incremental NDL model without its extension to account for trial-to-trial learning effects in lexical decision data of one subject in the BLP, showing that NDL models which reinforce their mappings after each trial show a better fit to speaker data than those without reinforcements. In the current study we explore a different implementation of discriminative learning in which form and meaning are coupled directly. Studies have pointed out the significance of semantics not only in lexical access and processing in general, but crucially also in the lexical decision task. Several studies found that variables related to a word's semantics, such as the semantic density of a word (Hendrix and Sun, 2021), its imageability (Balota et al., 2004), its availability of meaning (Chumbley and Balota, 1984) and how well its form predicts its meaning (Marelli et al., 2015; Marelli and Amenta, 2018; Hendrix and Sun, 2021) are predictive for reaction times in lexical decision.

Thus, we aim to model lexical decision with the DLM which represents words' semantics in a distributed fashion, using word embeddings, which have been found to model a remarkable number of phenomena in cognitive science (Günther et al., 2019). This setup has been successful at predicting a range of behavioural data related to lexical processing (e.g. Chuang et al., 2020b; Stein and Plag, 2021; Schmitz et al., 2021; Chuang et al., 2022a; Gahl and Baayen, 2022; Cassani et al., 2019; Heitmeier and Baayen, 2020). To model trial-to-trial learning we make use of the Widrow-Hoff learning rule (Widrow and Hoff, 1960) which allows learning of real-valued semantic vectors.

3 Introduction to the Discriminative Lexicon Model

The DLM is a model of the mental lexicon according to which comprehension is modelled as a mapping from form to meaning and production as a mapping from meaning to form. First, a way of representing word forms is needed. In this study, wordforms are transformed into binary cue vectors coding the presence and absence of trigrams in the wordform. As the present experiment addresses visual word recognition, we used letter triplets rather than triphones or

acoustic representations derived from the audio signal (Shafaei-Bajestan et al., 2021)[1]. By way of example, consider the wordform *aback*. As a first step, its set of unique trigrams is extracted (`#ab, aba, bac, ack, ck#`), with `#` denoting word boundaries. In a second step, in a vector where each value stands for a possible trigram in the lexicon, the trigrams present in *aback* are now coded with 1, all others with 0. The resulting vector is stored in a matrix **C** together with the form vectors of all other wordforms in the lexicon:

$$\mathbf{C} = \begin{array}{c} \\ \text{aback} \\ \text{back} \\ \text{...} \\ \text{lack} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \texttt{\#ab} & \texttt{aba} & \texttt{bac} & \texttt{ack} & \texttt{ck\#} & \texttt{\#ba} & \texttt{\#la} & \texttt{...} & \texttt{lac} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & ... & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & ... & 1 \end{pmatrix} \end{array}.$$

A model of the mental lexicon also requires representations for words' meanings. Word meanings are likewise coded numerically, making use of high-dimensional distributed representations, the individual dimensions of which do not necessarily have a straightforward interpretation. Importantly, vectors which are more similar to each other are also closer in meaning. Many different kinds of semantic vectors have been used to represent meaning in DLMs. A comparison of simulated vectors with vectors generated using approaches from NLP, such as *Word2Vec* (Mikolov et al., 2013), can be found in Heitmeier et al. (2021). In the present work, we explored various sets of vectors created using *Word2Vec* (Mikolov et al., 2013) and *GloVe* (Pennington et al., 2014). We obtained the best model fit for Subject 1 using *Grounded GloVe* vectors (Shahmohammadi et al., 2021). These are created by aligning existing *GloVe* vectors with information from images, without letting the vectors deviate far from their original text embeddings. In this way, the vectors absorb some of the information available in images, but do not lose abstract information which is only available in text. The set of vectors found to perform best on various benchmark tests in Shahmohammadi et al. (2021) have a dimensionality of 1024, which we accordingly used in our simulations.

Words' semantic vectors are stored in a matrix **S** (values in the following example are simulated):

$$\mathbf{S} = \begin{array}{c} \\ \text{aback} \\ \text{back} \\ \text{...} \\ \text{lack} \end{array} \begin{array}{c} \begin{array}{ccccccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 & S_7 & ... & S_n \end{array} \\ \begin{pmatrix} -0.11 & 0.13 & -0.06 & -0.16 & -0.33 & 0.46 & 0.37 & ... & 0.13 \\ 0.22 & 0.32 & -0.28 & -0.42 & -0.19 & 0.37 & -0.24 & ... & 0.01 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ -0.11 & 0.4 & -0.02 & -0.21 & -0.31 & -0.09 & 0.34 & ... & -0.16 \end{pmatrix} \end{array}.$$

For modeling comprehension, we use a mapping **F** that approximates **S** from **C**. As it is not a perfect mapping, we write

$$\hat{\mathbf{S}} = \mathbf{CF}. \tag{1}$$

For any individual word form (represented as a binary vector $\mathbf{c}$), we can obtain its meaning (predicted semantic vector $\hat{\mathbf{s}}$) via

$$\hat{\mathbf{s}} = \mathbf{c} \cdot \mathbf{F}. \tag{2}$$

In the same way we can also model production as a mapping from a word's semantics to its form. This is achieved simply by a mapping in the opposite direction, so from **S** to **C**, using the

[1]Many other representations are possible, such as features for orthographic input based on Histograms of Oriented Gradient features (Dalal and Triggs, 2005; Linke et al., 2017) (further overview in Heitmeier et al., 2021).

mapping matrix $\mathbf{G}$. Again, this mapping is not perfectly accurate:

$$\hat{\mathbf{C}} = \mathbf{SG}. \tag{3}$$

$\mathbf{G}$ can now likewise be used to obtain a word's predicted form ($\hat{\mathbf{c}}$) from its meaning ($\mathbf{s}$):

$$\hat{\mathbf{c}} = \mathbf{s} \cdot \mathbf{G}. \tag{4}$$

There are two ways in which $\mathbf{F}$ and $\mathbf{G}$ can be computed. The first is the so-called endstate-of-learning. Here, the mapping matrices are optimal, i.e. all words in the mental lexicon are learned as perfectly as possible. The mathematical engine behind this process is equivalent to multivariate multiple regression. Details on how the endstate-of-learning can be estimated efficiently can be found in Baayen et al. (2018) and Luo (2021).

The second option is to learn the mappings incrementally. Here the mappings are updated each time a word is encountered. As expected, the mapping between a word's form and its meaning in the mental lexicon becomes more accurate the more often it is encountered. Mathematically, this is achieved with the Widrow-Hoff learning rule (Widrow and Hoff, 1960). First, focus on word comprehension. When at time step t a word w_t is encountered, which has a word form $\mathbf{c}_t$ and a meaning $\mathbf{s}_t$, the mapping from form to meaning is updated, in a way which decreases the error between the predicted and the target semantics, making the learning "error-driven":

$$\mathbf{F}_{t+1} = \mathbf{F}_t + \mathbf{c}_t^T \cdot (\mathbf{s}_t - \hat{\mathbf{s}}_t) \cdot \eta_1 \tag{5}$$

Since the next time the same word is encountered, the mapping will be more accurate, we refer to this step as "strengthening" of the mapping for the remainder of this paper. There is one hyperparameter, η_1, which represents the learning rate. A higher learning rate implies not only that a form-meaning association is learned faster, but also that form-meaning associations which are not encountered are unlearned faster. Secondly, for production the same equation is used to update the $\mathbf{G}$ matrix:

$$\mathbf{G}_{t+1} = \mathbf{G}_t + \mathbf{s}_t^T \cdot (\mathbf{c}_t - \hat{\mathbf{c}}_t) \cdot \eta_1 \tag{6}$$

Finally, having computed the predicted form and semantic vectors, we need a way to evaluate how accurate the mappings are. For comprehension, the goal is that if a word form is encountered, the retrieved meaning, i.e. its predicted semantic vector, is as close to the target meaning of the word as possible. To evaluate how well this goal is approximated we compute the correlation between the predicted semantic vector $\hat{\mathbf{s}}_w$ of word w and all semantic vectors in $\mathbf{S}$. If $\hat{\mathbf{s}}_w$ has the highest correlation with its target semantic vector $\mathbf{s}_w$, the predicted meaning is closest to the target meaning, and therefore, it is counted as correct. We call this measure of accuracy in our model "correlation accuracy". Note that when mapping from form to semantics, an approximation of the semantic vectors is straightforwardly obtained. In production, on the other hand, the result of the mapping is a vector with varying support for various trigrams. In order to actually produce a word, it has to be decided a) which trigrams have enough support to be included in the produced wordform and b) in which order the trigrams should be produced. Since the trigrams are partially overlapping, they include information about possible orderings. Various algorithms for solving this problem are feasible, but as the production algorithm is not relevant for our current study, we point the interested reader to different implementations described in Baayen et al. (2018) and Luo (2021), and note here only that the used algorithm is a kind of beam-search algorithm. Production can then be simply evaluated by comparing the produced form with the target word form.

4 Lexical decision with the Discriminative Lexicon Model

Similar to previous work both in discriminative learning models (Milin et al., 2017b; Baayen et al., 2013) and also other computational models such as the interactive activation model of

Dijkstra and Van Heuven (2002) we view lexical decision as a two-step process. First, the incoming stimulus is processed in the mental lexicon. Next, the decision is made by distinct cognitive control processes (as e.g. proposed by Redgrave et al., 1999; Gurney et al., 2001). Our focus in the present study lies on the trial-to-trial learning effects arising in the mental lexicon rather than on the decision mechanism.

Our analysis therefore proceeds in two steps: first, we use the DLM to model how incoming forms activate their meanings, updating the networks/mapping matrices at each trial. Then we extract various measures from the networks, designed to quantify factors that may affect decision making, such as the density of a word's semantic neighbourhood. They allow us to detect the fine effects of trial-to-trial learning, and will be introduced in detail below. In the second step the extracted measures are used to predict reaction times using nonlinear regression modeling.

The two steps are outlined in the following sections. First, a detailed account is given of how we model trial-to-trial learning in lexical decision with the DLM, followed by a description of the measures we calculate, at each trial, on the basis of the current state of the network. Sections 5 and 6 will describe how we use these measures to predict the time it takes to execute lexicality decisions.

4.1 Processing in the mental lexicon

In order to simulate the lexical knowledge that participants have before the experiment, we set up mappings between form and meaning for all the words that are encountered during the experiment. As described in Section 3, the DLM can learn words in two ways: One is to initialise the model with the endstate of learning, where we assume that the model has learned all words equally well. Another is to learn the vocabulary incrementally according to frequencies obtained from a corpus. Although the latter option is ideally more realistic, in the absence of data that are properly chronologically ordered so that developing word embeddings can be calculated, and that enable developing mappings from forms to these embeddings, we decided to initialize participants' lexicons using the endstate-of-learning.

We therefore initialise the mapping matrices $\mathbf{F}$ and $\mathbf{G}$ with the endstate-of-learning calculated for the entire set of 28,456 words in the BLP for which semantic vectors were available (details in Section 5.1). This results in an accuracy of 61% for comprehension (for 81% of the words their target semantics was among the five closest semantic neighbours). 50% of wordforms are produced correctly (65% correct forms among top 10 candidates). One reason for this relatively low production accuracy is presumably that we did not implement the second route via phonology proposed in the original implementation of the DLM (Baayen et al., 2019).

Having set up the model, we now go through the lexical decision experiment trial by trial (Figure 1 provides an overview of the different modeling steps that unfold at each subsequent trial). When encountering a letter string in trial t, first, its meaning is retrieved. Therefore, we turn the letter string into a binary cue vector $\mathbf{c}_t$ (step (A) in Figure 1) and use the comprehension mapping $\mathbf{F}_t$ to obtain the string's predicted semantic vector $\hat{\mathbf{s}}_t$ (step (B)), by computing

$$\hat{\mathbf{s}}_t = \mathbf{c}_t \cdot \mathbf{F}_t. \tag{7}$$

We assume that before the experiment, the participant does not generally distinguish words and nonwords, but builds this knowledge up during the course of the experiment.[2] We therefore create an additional mapping $\mathbf{D}$ which maps the cue vector $\mathbf{c}_t$ directly to a word/nonword outcome d_t (C) (see also Linke et al., 2017, for a similar setup for modelling lexical decision in

[2]This is different to e.g. Norris (2006)'s Bayesian Reader which assumes that speakers make word/nonword decisions not only during lexical decision, but whenever they read. We do not agree with this view and instead posit that speakers generally assume that letter strings they encounter are words (which they might not know the meaning of). Therefore, lexical decision is a metalinguistic task, which has to be learned during the experiment (a view that is supported by the fact that there are training trials in the BLP).

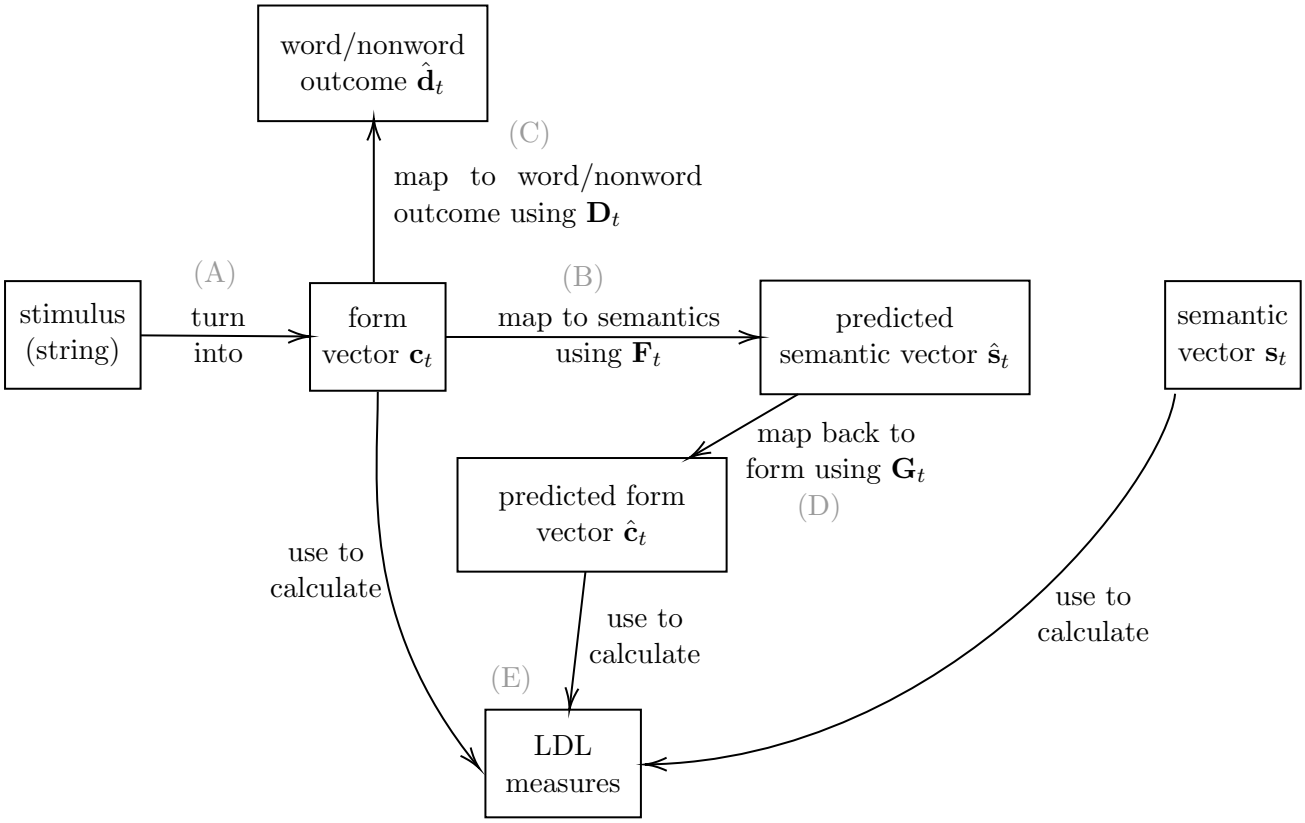


Figure 1: Overview over steps during simulation of one trial t. Boxes represent representations, while arrows show processes. The last step (not pictured) is to update $\mathbf{F}_t$, $\mathbf{G}_t$ and $\mathbf{D}_t$ using the Widrow-Hoff learning rule.

baboons). This matrix is initialised with zeros at the beginning of the simulation. By having a mapping matrix specialised in computing the word/nonword outcome, it is possible to have a different learning rate for this outcome than for the semantic outcomes. To obtain the support for the word/nonword outcome d_t given the cue vector $\mathbf{c}_t$ we compute

$$d_t = \mathbf{c}_t \cdot \mathbf{D}_t \tag{8}$$

Note that we do not view this network as a decision mechanism — rather, we assume that this bottom-up support for a word/nonword is one source of evidence for the decision mechanism also taking into account other processing measures (see below).

Next, we model a "feedback loop" in word comprehension (Chuang et al., 2020b). Multiple theories have suggested that speech production is involved in speech perception (e.g. Liberman and Mattingly, 1985; Skipper et al., 2017), based on empirical evidence such as neuroimaging studies (e.g. Pulvermüller et al., 2006). Moreover, phenomena such as "inner speech" suggest that feedback loops to production exist also during silent reading (e.g. Kell et al., 2017). To model this feedback loop, the predicted semantic vector[3] $\hat{\mathbf{s}}_t$ is mapped back to the form side to obtain the predicted form vector $\hat{\mathbf{c}}_t$ using the production mapping $\mathbf{G}_t$ (D):

$$\hat{\mathbf{c}}_t = \hat{\mathbf{s}}_t \cdot \mathbf{G}_t \tag{9}$$

Once we have obtained the predicted semantic and form vectors $\hat{\mathbf{s}}$ and $\hat{\mathbf{c}}$ we can proceed to calculate various measures (E) which can later be used to predict reaction times. These measures will be introduced and discussed in Section 4.2.

Finally, we model the learning which we hypothesise to take place after each trial. The participant's response is used to update all mappings (not displayed in Figure 1). Using the Widrow-Hoff learning rule (see equations 5 and 6 above), the mapping $\mathbf{F}_t$ from cue vector $\mathbf{c}_t$ to its target semantic vector $\mathbf{s}_t$ is updated, as well as the mapping $\mathbf{G}_t$ from $\hat{\mathbf{s}}_t$ to $\mathbf{c}_t$, both with learning rate $\eta_1 = 0.001$, which we found to give best results with our training subjects. This step predicts effects such as priming and antipriming as described in the introduction: if the same stimulus would be presented again, the mapping to its semantics would be more accurate than before the update. If a similar stimulus with very different semantics would be presented next, the mapping would be quite inaccurate, as the cues within the stimulus are mapped more strongly towards another meaning.

The target semantic vector for updating $\mathbf{F}$ necessarily depends on the response of the participant, since there was no feedback given to participants after each trial in the BLP. We distinguish four cases, depending on the participant's response (word/nonword) and the actual lexicality (word/nonword) (see Table 1). For word responses (to words), $\mathbf{s}_t$ is simply the semantic vector of word w_t in the semantic matrix $\mathbf{S}$. However, in trials where the participant responds with "word" but the string is actually a nonword, no target semantic vector for a nonword letter string exists. We therefore use the average of all word vectors in the participant's lexicon which we use to represent an average word meaning.

For nonword responses, we need a semantic representation for the concept of a "nonword". Without a nonword concept, it is not possible to update the mappings. We assume that participants do not have a concept of what a nonword is before the beginning of the experiment, but that this concept comes into being over the course of the experiment. We therefore calculate the nonword vector dynamically from previously encountered predicted nonword semantic vectors $\hat{\mathbf{s}}_{t_{nw}}$ where t_{nw} are trials with nonword responses. We compute the target vector for nonwords (see Supplementary Materials for alternative implementations) $\mathbf{n}_{t_{nw}}$ (used as target in the next trial with a nonword response $t_{nw}+1$) as

$$\mathbf{n}_{t_{nw}} = \frac{(\mathbf{n}_{t_{nw}-1} + \hat{\mathbf{s}}_{t_{nw}})}{2} \tag{10}$$

[3]Note that this definition differs from the usual definition of $\hat{\mathbf{c}}_t$ (found in e.g. Baayen et al., 2018) in that it is derived from the predicted semantic vector $\hat{\mathbf{s}}_t$ instead of the target semantic vector $\mathbf{s}_t$ since we do not assume that the target semantic vector $\mathbf{s}_t$ is available to participants at this point of processing.

Lexicality	Response = Word	Response = Nonword
Word	reinforce using word's semantic vector	reinforce using nonword vector
Nonword	reinforce using average of all semantic vectors	reinforce using nonword vector

Table 1: Decision table of which vector is chosen as target semantic vector for updating $\mathbf{F}$ after a trial.

where t_{nw} are trials with nonword responses. This implies that the concept of nonword is to 50% determined by the last stimulus with a nonword response, with the nonword encountered before that (according to the participant's response) contributing 25% of the vector, and so on. As a consequence, the nonword vector changes continuously, with the magnitude of change determined primarily by the nonword and its estimated semantic vector encountered previously.

The mapping matrix $\mathbf{D}_t$ is also updated using the Widrow-Hoff learning rule. Here, the target outcome is the participant's word/nonword response $r_t \in \{1, 0\}$. Note that $\mathbf{D}_t$ is not updated according to the actual lexicality of the string, but according to the participant's response. Since there is no "correct/incorrect" feedback in the BLP we thus model the participant's experience of the experiment:

$$\mathbf{D}_{t+1} = \mathbf{D}_t + \mathbf{c}_t^T \cdot (r_t - d_t) \cdot \eta_2 \tag{11}$$

with $d_t = \mathbf{c}_t \cdot \mathbf{D}_t$. We found that for Subject 1 setting the learning rate to $\eta_2 = 0.01$ yielded the best results. This suggests that the learning rate for the word/nonword outcome is higher than the learning rate used to reinforce the mappings between forms and meanings. This seems reasonable, as the lexical decision task requires subjects to make metalinguistic judgements, a cognitive task that subjects do not have much experience with, and that they learn to optimize as the experiment unfolds (Baayen et al., 2022). By contrast, one would expect lexical knowledge in long-term memory to be much less affected by trial-to-trial contingencies. In what follows, we used the same learning rates $\eta_1 = 0.001$ and $\eta_2 = 0.01$ for all subjects.

4.2 Predicting reaction times

In order to detect effects of trial-to-trial learning we need to quantify processing in the DLM. In the following, we will describe a few classical psycholinguistic measures, which are traditionally employed to predict lexical decision reaction times, as well as the measures we gauge from the DLM.

4.2.1 Classical predictors

We focus on three psycholinguistic, non-incremental predictors which have been used frequently to model lexical decision reaction times (e.g. Balota et al., 2004; Keuleers et al., 2012; Yap et al., 2015):

- **Word Frequency** Word frequency is generally associated with shorter reaction times in lexical decision tasks (e.g. Rubenstein et al., 1970; Scarborough et al., 1977), a finding replicated also for the BLP (Keuleers et al., 2012). We used word frequency in the British National Corpus[4], as reported in the BLP data. Though subtitle frequencies have been reported to be superior at predicting reaction times (Brysbaert and New, 2009), we opted for the BNC because it covers all registers and to avoid the confound of frequency and arousal found in subtitle corpora (cf. Baayen et al., 2016).

- **Word length** The effect of word length (i.e. number of letters) has been found to be somewhat inconsistent with some studies reporting null effects and others clear, negative

[4] `http://www.natcorp.ox.ac.uk`

effects of word length (overview in New et al., 2006). Possibly, null effects arise from a failure to match word and nonword stimuli in lexical decision experiments, see Chumbley and Balota (1984). New et al. (2006) linked this to a u-shaped effect of length (see also Baayen, 2005). They found that in the ELP, word lengths up to 5 letters tend to give rise to shorter reaction times, and for lengths from 8 to 13 letters to longer reaction times. No effect was found for lengths between 5 and 8 letters. Hendrix and Sun (2021) found that the effect of length changes across the distribution of reaction times. Early responses are unlikely for long words presumably because of higher visual processing costs linked to longer words. For short words, early responses are much more likely. This switches for somewhat later responses: these are more likely for longer words. Very late responses are equally likely for all word lengths. For nonwords on the other hand, multiple studies found that word length elicits longer reaction times (Yap et al., 2015; Balota et al., 2004).

- **Neighbourhood Size** Orthographic neighbourhood size was not found to be predictive for word reaction times in various virtual experiments, where reaction times for stimuli used in other studies were retrieved from the BLP (Keuleers et al., 2012). Andrews (1992) and Balota et al. (2004) on the other hand reported that larger neighbourhood size elicited shorter reaction times. For nonwords, Yap et al. (2015) and Balota et al. (2004) observed that neighbourhood size led to longer reaction times. Similar to *Word length*, the effect of neighbourhood size thus seems to be somewhat unclear with regard to words, but clearly leads to longer reaction times for nonwords. In our present analysis, we quantified orthographic neighbourhood size by the number of words in CELEX (Baayen et al., 1995) with a Levenshtein distance (Levenshtein et al., 1966) of 1 from the target stimulus.

We added two further predictors more concerned with the task itself than with lexical processing:

- **Trial number** Participants generally not only adapt to the task and become faster (Keuleers et al., 2012) but also speed up and slow down with varying levels of attention (Baayen et al., 2017, 2022), thus the current trial number is highly predictive for reaction times.

- **Response** We also included the participant's response (word/nonword) as a binary predictor in the model. We know from previous work that responses to words and nonwords tend to differ systematically (Keuleers et al., 2012) (how precisely they differ depends e.g. on the type of nonword used in the experiment, see Ratcliff et al., 2004). Since both correct and incorrect responses are an integral part of the learning process, we included both types of responses and and therefore added a factorial predictor to our models representing response type.

4.2.2 Measures from the DLM

From the DLM, we derived five measures for predicting the reaction times in the BLP. Of all the measures that we investigated (see the Supplementary Materials[5] for a full listing), these five proved to be the best predictors for the response latencies of our training subjects.

The first two groups of measures are extracted from the core DLM, involving semantics and the comprehension and production matrices **F** and **G**. As these measures are directly related to lexical processing, we will refer to them as *lexical-DLM* measures. The first two measures address words' semantic and orthographic neighborhoods.

- **Semantic Density** A word's semantic density, i.e. the number and proximity of semantic neighbours, has been used in previous work to predict not only reaction times in lexical

[5]Supplementary Materials including the simulation code, all generated measures and statistical analyses can be found at `https://osf.io/bxmt2/`.

decision (e.g. Hendrix and Sun, 2021; Buchanan et al., 2001; Chuang et al., 2020b; Schmitz et al., 2021; Stein and Plag, 2021), but also in other fields such as word learning (Hopman et al., 2018). The measure of semantic density that we have found to be optimal is based on the closest semantic neighbors of the predicted semantic vector $\hat{s}$, and gauges how densely populated the area around the predicted semantic vector $\hat{\mathbf{s}}$ is in semantic space. If a form lands in a semantically dense area, this indicates not only that it has presumably landed in an area of high lexicality, i.e. "wordlikeness", but also that it might be more difficult to tell the meaning of the word apart from similar meanings (Arnold et al., 2017). Semantic density can be quantified by inspecting the n closest semantic neighbours and computing the mean of their cosine similarities to $\hat{\mathbf{s}}$ (see e.g. Buchanan et al., 2001). Let CS_t be the set of all cosine similarities between $\hat{\mathbf{s}}_t$ and the semantic vectors $\mathbf{s}_k \in \mathbf{S}$:

$$CS_t = \{\text{cosine\_similarity}(\hat{\mathbf{s}}_t, \mathbf{s}_k) \,\forall\, \mathbf{s}_k \in \mathbf{S}\} \tag{12}$$

Then, *Semantic Density* is defined as the mean of the n highest values in CS_t:

$$\text{Semantic Density}_t = \frac{\sum \max_n(CS_t)}{n} \tag{13}$$

We set $n = 10$.

- **Shortest Path** *Shortest Path* combines insights from orthographic and semantic neighbourhood measures. It is motivated by two findings from previous work. First, we know from studies such as Forster and Davis (1984); Rodd (2004); Bowers et al. (2005) that during word recognition, not only orthographic neighbours but also subsets and supersets of words (including their semantics) are activated. Secondly, Marelli et al. (2015) found that a measure of the semantic similarity between a word's orthographic neighbours (Orthographic-Semantic Consistency, OSC) is predictive for lexical decision reaction times in the BLP. *Shortest Path* aims to account for both of these findings by quantifying how far apart the orthographic neighbours of a stimulus are in semantic space.

  We take the set of nearest neighbours N (we approximated this with the *Coltheart's N* neighbours, i.e. all orthographic word neighbours (in the BLP) of the same length with one letter exchanged), and calculate the corresponding predicted semantic vectors $\hat{\mathbf{s}}_n$ for $n \in N$. Then we find the shortest path in semantic space (measured in Euclidean distance) connecting all predicted semantic vectors $\hat{\mathbf{s}}_n$ including the predicted semantic vector of the target stimulus $\hat{\mathbf{s}}_t$. Finding the shortest path is a case of the *Travelling Salesman Problem*, where the goal is to find the shortest path connecting all points in a multi-dimensional space (see Figure 2). We made use of algorithms by Pferschy and Staněk (2017), implemented in Julia[6].

  For the 54% of words in the BLP for which OSC is available in Marelli and Amenta (2018), correlation between *Log Shortest Path* and OSC is $r = -.34$.

The second set of measures quantifies the "feedback loop", i.e. the production mapping from a stimulus' semantics $\hat{\mathbf{s}}_t$ back to its form ($\hat{\mathbf{c}}_t$) via the production mapping $\mathbf{G}_t$.

- **C-Precision** *C-Precision* is a measure of how well the predicted form vector $\hat{\mathbf{c}}_t$ matches the original form vector $\mathbf{c}_t$ representing the seen letter string. As such, it measures how precise the mapping back from the predicted semantics ($\hat{\mathbf{s}}_t$) to the form level is: the more similar the two are, the higher the *C-Precision*. Mathematically, it is defined as:

$$\text{C-Precision}_t = cor(\mathbf{c}_t, \hat{\mathbf{c}}_t) \tag{14}$$

[6] `https://github.com/ericphanson/TravelingSalesmanExact.jl`

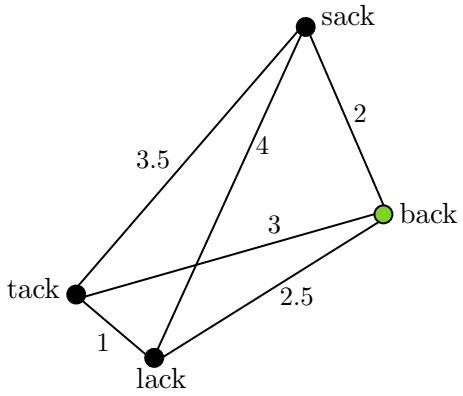


Figure 2: Four points in a two-dimensional semantic space, with hypothetical (euclidean) distances between them. The green node is the vector $\hat{\mathbf{s}}$ for the target word *back*, the others represent the semantic vectors of four of its orthographical neighbours. *Shortest Path* measures the shortest path connecting all points. In this toy example, the shortest path would be *back* → *lack* → *tack* → *sack* → *back*, with a length of 9. The *Shortest Path* for this example therefore is 9.

- **L1Chat** *L1Chat* measures the uncertainty in the predicted form vector $\hat{\mathbf{c}}$ (similar to the activation diversity measure in NDL, see e.g. Milin et al., 2017b). If *L1Chat* is large, there is a lot of support for many cues, either in a relatively distributed fashion across many cues, or centered on a few, highly supported cues. It is defined as

$$\text{L1Chat}_t = \sum_{j=1}^{n} |\hat{c}_j| = L_1(\hat{\mathbf{c}}_t) \tag{15}$$

  with n the length of the vector $\hat{\mathbf{c}}$.

The last measure is designed to measure the "wordlikeness" of a word form. It is derived from the network **D** introduced specifically for the modeling of lexical decision making. Therefore, it is not directly related to lexical processing itself, but rather a task-specific, learned measure. It is therefore not included in the set of *lexical-DLM* measures.

- **Yes-activation** *Yes-activation* is simply the support for the outcome "Word" in the decision vector d_i (see Section 4.1). It thus measures how strongly the sublexical cues of a stimulus support a word outcome given the participant's previous experience with words. This measure is available only to simulations in which networks are updated from trial to trial (henceforth 'dynamic simulations').

5 Data preprocessing and analysis

We used the measures described in the previous section to predict reaction times for all subjects using Generalised Additive Models (GAMs; Hastie and Tibshirani, 1987), as implemented in the *mgcv* package (Wood, 2011) for R. GAMs are regression models that can incorporate non-linear effects of one or more predictors on the response variable (see also Baayen et al., 2022). The following sections describe data preprocessing, selection criteria for the measures we used in the final models and details on our regression modelling strategy.

5.1 Data

We used the data collected by Keuleers et al. (2012) in the British Lexicon Project (BLP). They collected lexical decision reaction times for 28,730 words and an equal number of nonwords from

78 British students. To save time — the experiment took about 16 hours per participant —, each participant responded to half of the stimuli. Words with a frequency of at least 0.02 per million were selected. The nonwords were generated using Wuggy (Keuleers and Brysbaert, 2010), making use of the following criteria: the word and nonword matched in syllabic and subsyllabic as well as in morphological structure, monosyllabic nonwords differed in one and disyllabic ones in two subsyllabic elements from the base word and transition frequencies of subsyllabic elements were matched as much as possible. As described in previous work, even though all nonwords were based on real words, the method used to generate them made most nonwords opaque as to their base words (Hendrix and Sun, 2021).

Selecting all words in the BLP which could be found in the grounded *GloVe* vectors (Shahmohammadi et al., 2021) resulted in a set of 28,465 words. Before the simulation, we removed trials with 'null' and 'nan' as target stimuli (156 datapoints), as these spellings disrupted data processing, and removed all trials with time-out responses since they did not have a clear word/nonword response (21 responses for subject 65, 4 for subject 70 and 1 for subject 10).

For data analysis, we excluded all trials with reaction times $\leq$ 100ms, which is the minimum for response execution, or $>$ 2000ms, which are outliers in the distribution and probably reflect additional cognitive processes which are not of interest to the present study (removing 20,094 trials (0.9%)). Reaction times were transformed in the following way:

$$RTinv = -1000/RT \tag{16}$$

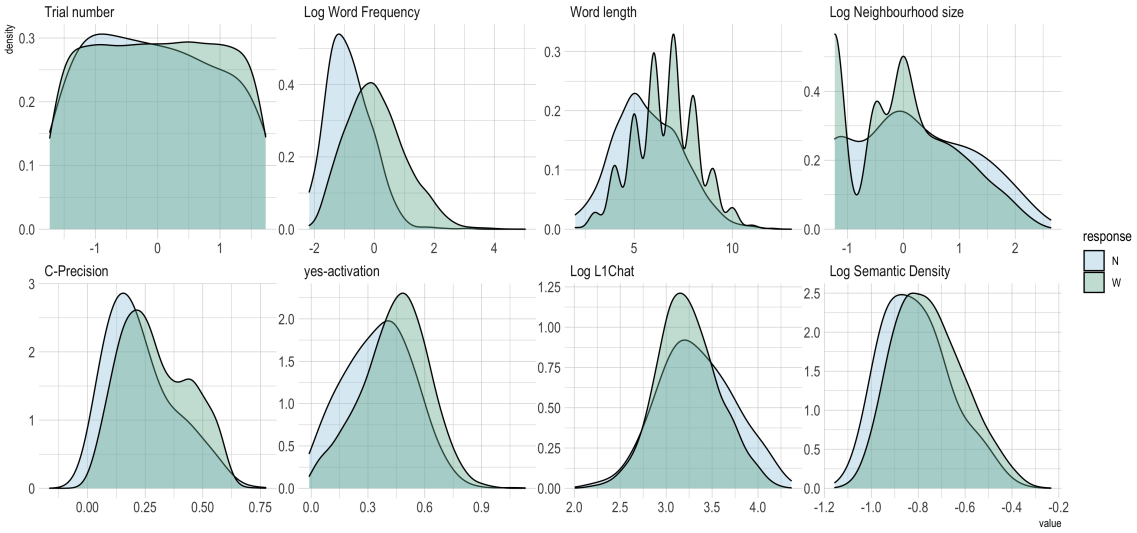
to obtain an approximately normal distribution for the response variable, suitable for analysis with Gaussian GAMs. This transformation implies that instead of response time, we study response rate (with a scaling factor 1000 to avoid very small numbers, and negative sign to ensure a positive correlation of the rate variable with the time variable). However, for ease of reading we will refer to this negative response rate as "reaction time" for the remainder of this paper.

The distribution of each predictor variable was checked and log transformed after adding a backup value of 0.002 if necessary to reduce outlier effects (distributions for subject 1 are available in Figure 3). If a predictor p has a substantial number of zeros in the original distribution, log transformation can lead to a bimodal distribution (see e.g. *Log Neighbourhood size* in Figure 3a). For such predictors, we introduced a binary variable b indicating whether the untransformed predictor was zero. When adding the predictor p to the GAM, we additionally specified the binary predictor b and the $p \times b$ interaction term to avoid that a large part of the probability mass of a predictor is concentrated in a single (typically extreme) value. This procedure was applied to *Log Word frequency* (binary predictor *in_bnc*), *Log Neighbourhood size* (binary predictor *has_neighbours*) and *Log Shortest Path* (binary predictor *has_neighbours_path*). *Trial number* was centered and scaled. This procedure was applied to the data of all subjects.

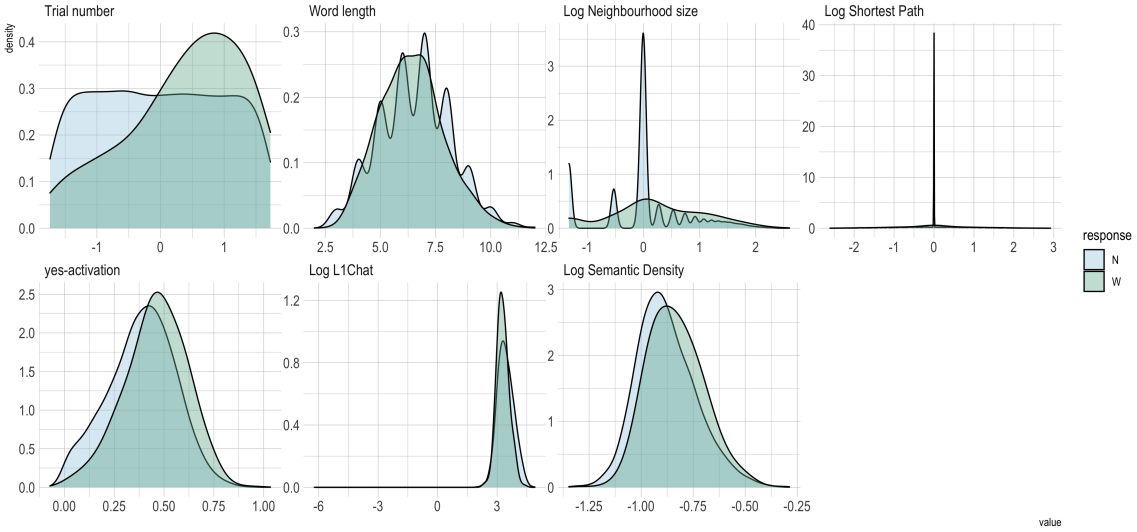
5.2 Predictor selection

During initial exploratory data analysis using only the data from the two training subjects, we experimented with a set of measures obtained from the model (full list in the Supplementary Materials) and used these to predict reaction times. Predictors were included in our exploratory models if, and only if, (1) their partial effect was significant ($p < 0.001$), (2) including the predictor improved the overall Akaike Information Criterion (AIC; Akaike, 1998), and (3) inclusion of a predictor did not lead to unacceptably high concurvity[7]. Correlation between all DLM-based predictors was $< .6$. As detailed in the Supplementary Materials, there were two exceptions to

[7] Concurvity, the counterpart of collinearity in the strictly linear model, estimates the extent to which the partial effect of a given predictor can be accounted for by the other predictors in the model. When concurvity is high, it is unclear whether predictors with high concurvity scores make an independent contribution to the model fit. For discussions of collinearity and concurvity, see Friedman and Wall (2005) and Tomaschek et al. (2018). Further details on the exploratory modeling are provided in the Supplementary Materials.



(a) Words



(b) Nonwords

Figure 3: Distribution of measures for words and nonwords for subject 1. To make the distributions more visible, the area under each density plot sums to 1, and therefore does not reflect the true proportions of word and nonword responses.

these rules: *C-Precision* in the word models and *Yes-activation* in the nonword models did not reach significance in one of two training subjects, but did substantially improve model fit, and were therefore included.

5.3 Regression modeling strategies

The BLP dataset is too large to allow fitting with an insightful generalized additive mixed model. To avoid this computational bottleneck, we fitted separate GAMs to the data of the individual subjects. Furthermore, for ease of interpretation, we fitted separate models to the word data and to the nonword data.

The sequences of reaction times in the BLP form time series that are characterized by autocorrelations (e.g. Baayen et al., 2017). GAMs can take autocorrelations into account by building an AR(1) process into the residuals, such that the residual at t is a proportion ρ of the residual at $t-1$ plus Gaussian noise. We obtained ρ for each model individually by first extracting ρ from a GAM without autocorrelation with classical predictors for both words and nonwords respectively. We then set this value as our ρ for the subject, and ran both classical and DLM-based models, this time with autocorrelation. As the original BLP experiment was too long to perform all in one session, the participants were allowed to freely choose how many blocks they wanted to do in one day. A session expired after a break of more than 10 mins between blocks. Since we assumed that after such a break, a response would not be influenced by the previous one anymore, we opted to restart the autocorrelation for each new session. We experimented with never restarting and restarting only for each new day of the experiment, but found that a session-based restart provided the best fit for our training subjects.

Model criticism revealed that the de-correlated residuals did not follow Gaussian distributions. As a consequence, our models remain approximate. To ensure that these approximate models are reasonable, we also considered Gaussian location-scale models, which model the effect of predictor variables on the dependent variables median and variance, as well as Quantile GAMs, which are distribution free. The functional form of partial effects remained stable across these analyses. Full details are available in the Supplementary Materials.

We complemented the GAM analyses (Sections 6.1 and 6.2) with LMMs fitted to the data of all subjects jointly, with one LMM fitted to the word data, and one to the nonword data. These models are reported in Section 6.3.

6 Results

In what follows we first analyse the individual predictors, their effects and reliability across subjects. We then answer the main question of this study, namely whether trial-to-trial learning can be detected in the BLP data. Finally, we take a closer look at individual differences between subjects.

6.1 Modeling reaction times to words and nonwords with GAMs

6.1.1 Words

Baseline: Classical Predictors We started out by fitting a baseline model using only classical psycholinguistic measures (*Log Word frequency*, *Word length* and *Log Neighbourhood size*) to predict reaction times. This model cannot take trial-to-trial learning into account. Additionally, we included *Trial number* and the participant's *Response* (word/nonword) as predictors. In the following, we will refer to the ratio of subjects for which an effect is significant ($p < 0.001$) as a predictor's "reliability". An overview over the various predictors, the direction of their effect and reliability can be found in Table 2.

Predictor	Increase elicits...	Reliability
Trial number	shorter RTs (but wiggly)	100%
Log Word frequency	shorter RTs, attenuated at high frequencies	100%
Word length	longer RTs	87%
Log Neighbourhood size	65% longer RTs, 33% shorter RTs, rest U-shaped	55%
Response=W	30% shorter RTs, 70% longer RTs	82%

Table 2: Predictors and their reliability for **words** in the **classical GAMs**. Effect of increase (given for significant predictors only) is intended as a summary and may differ for individual subjects (see Figure 4 for details). Reliability gives the percentage of subjects for which the predictor (regardless of direction) is significant ($p < 0.001$).

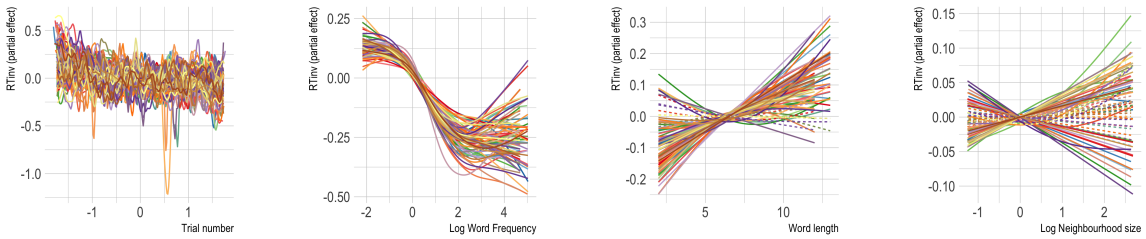


Figure 4: Partial effects of **classical predictors** for all subjects (**words**). Solid lines have a significance level of $p < 0.001$. While the effects of *Log Word frequency* is very similar across all subjects, the effects of *Word length* and *Log Neighbourhood size* show variability, indicative of individual differences.

Trial number was a significant predictor for all subjects (see Table 2). Inspecting the individual effects, we see that along the course of the experiment (see Figure 4), reaction times generally became shorter, with a couple of exceptional subjects who remained relatively stable and others who even slowed down. There was also considerable variability within sessions (cf. Pham and Baayen, 2015; Baayen et al., 2017). *Response* was significant for 82% of participants. The third predictor, *Log Word frequency*, was also significant for all subjects. The effect was qualitatively remarkably similar across all subjects. Higher *Log Word frequency* generally elicited shorter reaction times. At very high frequencies this effect was attenuated (Baayen et al., 2006; Keuleers et al., 2012). Higher *Word length* (significant predictor for 87% of subjects) gave rise to longer reaction times, except for five subjects for which the effect was U-shaped. The U-shaped effect reported by Baayen (2005) and New et al. (2006) apparently did not generalize to the majority of participants in the BLP. Finally, the most contested predictor for words, *Log Neighbourhood size*, was significant in 55% of cases. The direction of the effect was quite incoherent across subjects. For 28 of the subjects higher *Log Neighbourhood size* elicited longer reaction times, whereas for 14 they were shorter. The remaining subjects' effects either were not significant or had no clear direction (one subject). This variability is presumably one of the reasons why the effect of *Log Neighbourhood size* was found to be so inconsistent across previous studies (Keuleers et al., 2012; Andrews, 1992; Balota et al., 2004).

DLM measures We included two sets of measures in our GAMs: a set of non-incremental measures (*Trial number*, *Log Word frequency*, *Word length* and *Response*), and the incremental measures from the DLM (*Log Semantic density*, *Log L1Chat*, *C-Precision* and *Yes-activation*). An overview of all predictors as well as their reliability can be found in Table 3, all DLM-based effects are visualised in Figure 5. The classical predictors in the dynamic GAMs had similar effects as in the baseline model, and are therefore not displayed (but see Supplementary

Materials for further details).

We again included *Trial number*, since we assumed that effects which arise from e.g. increased motor training, task adaption or attention fluctuations (cf. Baayen et al., 2022) cannot be explained by our model. Next, we also included *Log Word frequency*. Our model can in principle also account for word frequency effects by initialising the comprehension and production matrix with incremental learning, where more frequent words are trained more often and thus learned better. However, this procedure is computationally very costly when using the Widrow-Hoff learning rule and thus currently unfeasible. To reduce the carbon footprint of our simulations, we therefore opted to initialise our models with the much more efficient "endstate-of-learning", which prevents the model from showing frequency effects, and instead included *Log Word frequency* as a predictor in the GAMs. The third non-incremental variable was *Word length*, a crude way of estimating the complexity of the eye-reading process (Engbert et al., 2002). We did not include *Log Neighbourhood size*. *Trial number* and *Log Word frequency* were significant for all subjects. *Word length* was a significant predictor for somewhat fewer subjects (74%), and *Response* for 77%.

The remaining predictors were grounded in the DLM. The first measure in the top left of Figure 5 is *Log Semantic density*. Its effect was significant for 56% of all subjects: the denser the semantic space the predicted vector $\hat{\mathbf{s}}$ landed in, the faster the reaction. This fits well with insights gained with models such as MROM, where higher general activation implies higher lexicality — and thus faster reaction times (Grainger and Jacobs, 1996).

The measure to the right of *Log Semantic density* in Figure 5 is *Log L1Chat*, a measure of the uncertainty in the $\hat{\mathbf{c}}$ vector. Here we obtained the best model fit when conditioning the effect on the response given by the subject. *Log L1Chat* was reliable for both word responses (78% of subjects) and nonword responses (90% of subjects). If the response was nonword, higher *Log L1Chat* was associated with shorter reaction times (i.e. high uncertainty led to faster reaction times for nonword responses), while for word responses it elicited longer reaction times (high uncertainty led to slower reaction times).

The first measure in the bottom row of Figure 5, *C-precision*, measures how correlated the predicted vector $\hat{\mathbf{c}}$ is with the original form vector $\mathbf{c}$. In other words, *C-precision* measures how precise the mapping back from the semantics to the form level is. It was significant for about half of the subjects. For these subjects, the more precise the mapping back to form was, the longer reaction times were. Our interpretation of this effect is that a well-supported form vector requires suppressing the production system more, which takes resources away from making a rapid lexicality decision.

The final measure was *Yes-activation*, which was significant for 32% of subjects. For these subjects, the more sublexical evidence in favour of a word outcome (higher *Yes-activation*) was available, the faster participants reacted.

We finally observed that 99% of the GAM models based on the DLM measures (with incremental updates) had a lower AIC value (i.e. better model fit) than the classical models (Mean/SD AIC difference 152.6/93.2; see also Figure 8).

6.1.2 Nonwords

Baseline: Classical predictors The nonword GAMs differed in some respects from the word models. As we had no frequencies for the nonwords in the BLP (but see Hendrix and Sun, 2021, for the predictivity of nonword frequencies from the web for lexical decision latencies), we only included *Trial number*, *Word length* and *Log Neighbourhood size* as well as *Response* as classical predictors in our baseline model. Their effects are visualised in Figure 6. *Trial number* was again eliciting shorter reaction times and *Word length* and *Log Neighbourhood size* both gave rise to longer reaction times, replicating results from previous studies (Yap et al., 2015; Balota et al., 2004). All three were significant for all subjects; response was significant for 82% of subjects (Table 4).

Predictor	Increase elicits...	Reliability
Trial number	shorter RTs (but wiggly)	100%
Log Word frequency	shorter RTs, attenuated at high frequencies	100%
Word length	longer RTs	74%
Log Semantic density	shorter RTs	56%
Log L1Chat (word response)	longer RTs	78%
Log L1Chat (nonword response)	shorter RTs	90%
C-Precision	longer RTs	51%
Yes-activation	shorter RTs	32%
Response=W	35% shorter RTs, 65% longer RTs	77%

Table 3: Predictors and their reliability for **words** in the **DLM-based models**. Effect of increase (given for significant predictors only) is intended as a summary and may differ for individual subjects (see Figure 5 for details). Reliability gives the percentage of subjects for which the predictor (regardless of direction) is significant ($p < 0.001$).

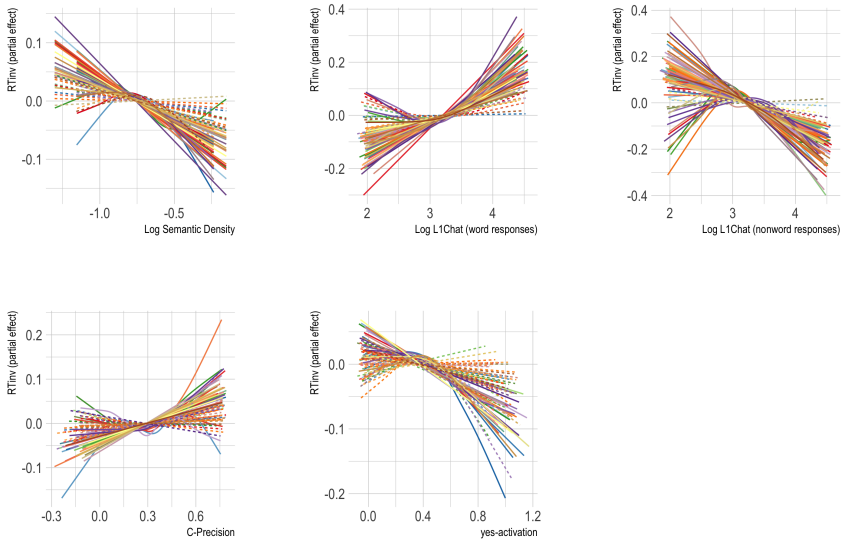


Figure 5: Partial effects of **DLM predictors** for all subjects (**words**). Solid lines have a significance level of $p < 0.001$. Classical measures are omitted, as their partial effects are very similar to Figure 4. The ranges of predictors vary within plots as a consequence of the between-subject design of the BLP. Full figures can be found in the Supplementary Material.

Predictor	Increase elicits...	Reliability
Trial number	shorter RTs (but wiggly)	100%
Word length	longer RTs	100%
Log Neighbourhood size	longer RTs	100%
Response=W	16% shorter RTs, 84% longer RTs	82%

Table 4: Predictors and their reliability for **nonwords** in the **classical GAMs**. Effect of increase (given for significant predictors only) is intended as a summary and may differ for individual subjects (see Figure 6 for details). Reliability gives the percentage of subjects for which the predictor (regardless of direction) is significant ($p < 0.001$).

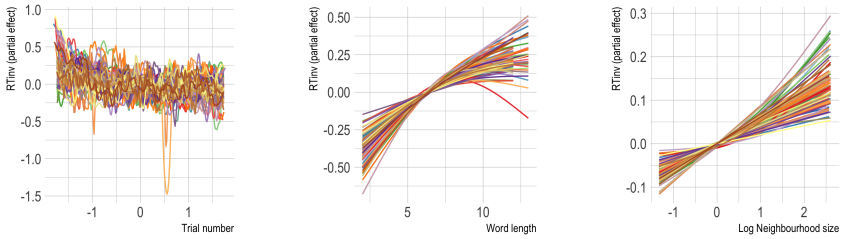


Figure 6: Partial effects of **classical predictors** for all subjects (**nonwords**). Solid lines have a significance level of $p < 0.001$. The effects are quite uniform across subjects.

DLM measures Turning again to the DLM measures, we included *Trial number* and *Word length* as non-incremental predictors for the same reasons as in the word models above; both were significant for nearly all subjects and had a remarkably similar effect as in the classical models (and are thus not displayed, but see Supplementary Materials). *Response* was significant for 88% of subjects. Additionally, we found *Log Shortest Path*, *Yes-activation*, and an interaction between *Log L1Chat* and *Log Semantic density* conditioned on response to be good predictors for nonword reaction times. All effects are summarised in Table 5 and visualised in Figure 7.

The left panel in Figure 7a shows the effect of *Log Shortest Path*, which was relatively reliable ($p < 0.001$ for 71% of subjects): The more orthographic neighbours of a nonword there were, and the further apart these neighbours were in semantic space, the longer it took a subject to react to the nonword. This is plausible because this effect dovetails well with the effect of *Log Neighbourhood size*, with which it is correlated (r = 0.67): the more orthographic neighbours a nonword has, the more it looks like a word, and the longer it takes to reject it as a word.

The next measure was *Yes-activation*. A higher *Yes-activation*, i.e. a higher support for a word outcome, predicted longer response latencies. As expected, its effect was opposite to its effect for words. While *Yes-activation* was only significant in 32% of subjects for words, it was significant for virtually all subjects for nonwords (99%).

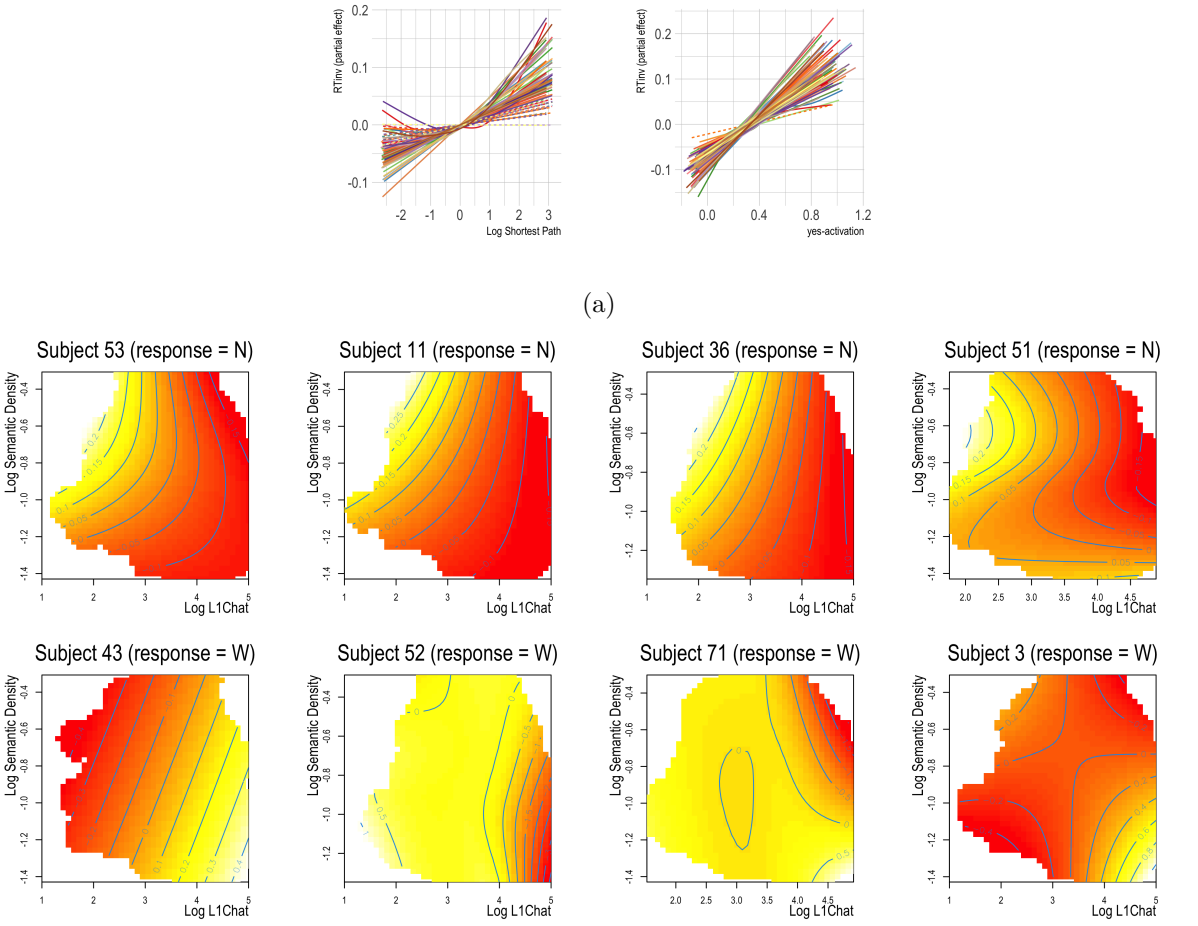
The interaction between *Log L1Chat* and *Log Semantic density* for nonword responses (modelled with tensor product smooths) was significant and quite uniform across all subjects. The left three panels in the upper row of Figure 7b, Subjects 53, 11 and 36 show a typical pattern: higher *Log L1Chat* elicited shorter reaction times, while higher *Log Semantic density* elicited longer reaction times for lower values of *Log L1Chat*. Subject 51 shows a somewhat wiggly effect of *Log Semantic density*. Similar plots for all subjects can be found in the Supplementary Materials.

For word responses the interaction between *Log L1Chat* and *Log Semantic density* was only significant for 22% of subjects. The patterns displayed by the effect differ significantly across subjects. The leftmost panel in the lower row of Figure 7b shows one of the most typical effect of the interaction between *Log L1Chat* and *Log Semantic density* for word responses. It can be seen that the relatively strong effect of *Log L1Chat* (eliciting longer reaction times) is reversed compared to nonword responses. The effect of *Log Semantic density* is not as pronounced. The other panels in the lower row of Figure 7b show subjects for which the effects were (partially) reversed and more interactive than for typical subjects.

Finally, we note that the GAMs based on DLM measures had a lower AIC (i.e. better model fit) than the classical models for all subjects (Mean/SD AIC difference 135.3/77.8; see also Figure 8). DLM measures seem therefore well suited to predict nonword reaction times.

Predictor	Increase elicits...	Reliability
Trial number	shorter RTs (but wiggly)	100%
Word length	longer RTs	99%
Shortest Path	longer RTs	71%
Yes-activation	longer RTs	99%
N response: Sem. Density x L1Chat	L1Chat shorter RTs, Sem. Density longer RTs (effect stronger for lower L1Chat)	100%
W response: Sem. Density x L1Chat	no generalisable effect in any direction	22%
Response=W	13% shorter RTs, 87% longer RTs	88%

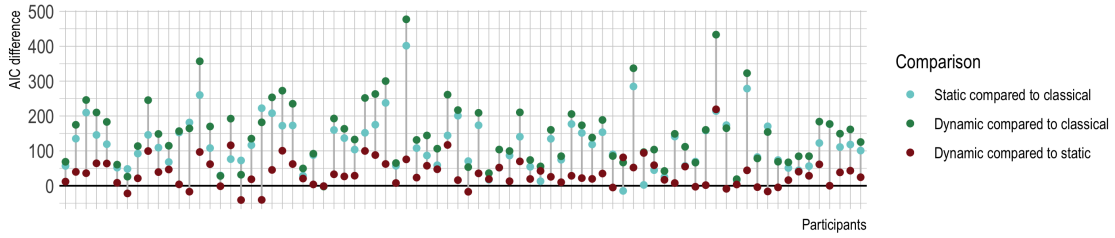
Table 5: Predictors and their reliability for **nonwords** in the **DLM-based GAM models**. Effect of increase (given for significant predictors only) is intended as a summary and may differ for individual subjects (see Figure 7 for details). Reliability gives the percentage of subjects for which the predictor (regardless of direction) is significant ($p < 0.001$).



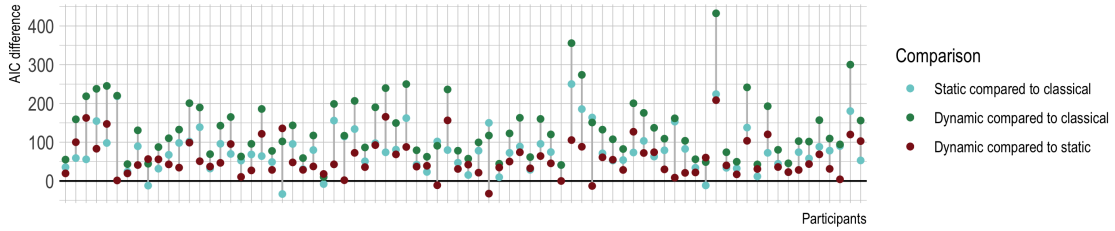
(a)

(b)

Figure 7: (a) Partial effects of the thin-plate regression smooth **DLM predictors** for all subjects (**nonwords**). Solid lines have a significance level of $p < 0.001$. Classical measures are omitted, as they are very similar to Figure 4. (b) Sample of tensor product partial effect for nonwords, yellow means longer, red shorter reaction times. Full figures can be found in the Supplementary Material.



(a) Words



(b) Nonwords

Figure 8: AIC comparisons of classical, static (i.e. no trial-to-trial learning) and dynamic (with trial-to-trial learning) models for both words (a) and nonwords (b). If, for example, the AIC difference of "static compared to classical" (turquoise) is positive for a subject, the static GAM has a better model fit than the classical one for this particular subject. The other comparisons can be interpreted analogously. Static and dynamic models almost always have higher relative likelihood than the classical model. Dynamic models mostly show a better model fit than static models, implying that the models benefit from trial-to-trial learning.

6.2 Trial-to-trial learning

In order to be able to answer the main question of this study, whether the modelling profits from incremental updates during the simulation, we ran two models for each subject. Both models included the same predictors, but were based on two different instances of the DLM: one with (*dynamic simulation*) and one without (*static simulation*) incremental learning. All other specifications of the two models were kept equal (with the exception of *Yes-activation*, which was only included in the *dynamic simulations*, see below). This allowed us to directly compare measures obtained from one model with and one without learning for each subject.

The dynamic models for words had lower AIC values than the static ones in 85% of cases. Difference in AIC values ranged from -40.9 (static better than dynamic) to 219.2 (dynamic better than static) (M 35.2, SD 40.8). On average, the relative likelihood of dynamic compared to static models was 5.0×10^{45}. For nonwords, dynamic models were better than static ones in 94% of cases (M 55.7, SD 45.8, ranging from -32.7 to 208.5), with the relative likelihood of dynamic compared to the static models on average 2.5×10^{43}. The differences in AIC values are presented in Figure 8.

Note that for static models we were unable to include the *Yes-activation* predictor, as it is critically dependent on incremental updates. This raised the question of whether the improved model fit of dynamic simulations was due to the incremental updates of the main mapping matrices $\mathbf{F}$ and $\mathbf{G}$ during the simulation, or whether it was mainly the *Yes-activation* that was responsible for improving goodness of fit. To investigate this possibility, we ran GAMs for the

dynamic simulations without *Yes-activation* (i.e. only classical and *lexical-DLM* measures) and again compared AIC values. We found that for word models even without *Yes-activation* dynamic GAMs still provided a better model fit for 82% of the subjects (M/SD AIC difference: 35.2/40.7). For nonwords, however, this was only the case for 60% (M/SD AIC difference: 3.0/24.7). Apparently, for responses to words, trial-to-trial updating of the lexical networks contributed substantially to the goodness of fit. However, for nonwords, improvements in goodness of fit were to a certain extent due to purely form-based sublexical learning.

6.3 Individual differences

As explained in the introduction, we expected individual differences to emerge in our simulations. While the individual GAMs reported in Section 6 show already considerable variability (see, for instance, Figure 4), individual differences can also be investigated with mixed models by including *participant* as a random effect. In principle, GAMs can include random effects. Unfortunately, for the large dataset of the BLP, we were confronted with two problems: a) the dataset is too big for the current implementation in *mgcv* to estimate a model with the full complexity that we need, and b) a Generalised Additive Mixed Model with all necessary interactions, even if it were estimable, would be extremely difficult to interpret. Therefore, to study individual variation within a regression framework, we needed to simplify. Fortunately, many partial effects in our GAMs are fairly linear (see Figures 5 and 7). We therefore ran a Linear Mixed Model (LMM) with all subjects with the same predictors as in the GAMs for words and nonwords respectively. We computed random by-participant intercepts as wells as random by-participant slopes for all predictors. We used the *julia* package *MixedModels.jl* (Bates et al., 2021) for efficient and precise model fitting.

Firstly, our two LMMs (see Table 6) for words and nonwords respectively confirmed the direction of effect for all predictors, and showed that all are significant ($p < 0.001$), even those with relatively low reliability in the individual GAMs such as *Yes-activation* for words. Exceptions were the main effects as well as the interaction of *Log Semantic density* and *Log L1Chat* for word responses in the nonword model ($p > 0.68$). Possibly, this was because only 5.7% (63,274) of responses to nonwords were word responses and because we were neglecting potential nonlinearities that cannot be modelled with the hyperbolic plane given by a standard multiplicative interaction in the linear modelling framework.

Since we do not have much meta-data about individual participants, we discuss individual differences in terms of their overall reaction time (fast vs. slow subjects). To understand subject-specific differences in the effects of our predictors, we make use of visualization by plotting by-subject random slopes against by-subject random intercepts (Figure 9).

We first consider individual variability as revealed by the three non-incremental, classical predictors in the mixed models: *Trial number*, *Log Word frequency* (in word models only) and *Word length*. The upper row of Figure 9a plots the correlations between participant-specific coefficients and the random intercept adjustment for the three non-incremental predictors (full correlation tables can be found in the Supplementary Materials). A negative intercept adjustment indicates fast participants, while a positive one reveals particularly slow subjects. The y-axis shows the participant-specific coefficients (random by-participant slope + overall coefficient). For values > 0, the overall effect of a predictor is positive, while for values < 0 it is negative. For *Trial number*, there is no solid correlation between intercept and random slope adjustment and the effect points in the same direction for most subjects, as most are below the zero-line. *Log Word frequency* elicits shorter reaction times for all subjects, as they are all below the zero-line, and its correlation with the random intercept adjustment is relatively weak (-.23). For *Word length* on the other hand, a clear relationship can be seen. For fast subjects (left side of the plot), the effect of *Word length* is weaker than for slow subjects (right side of the plot). The correlations within the nonword model are very similar and therefore not displayed here (see Supplementary Materials). We can therefore conclude that the non-incremental, classical predictors in our

	Est.	SE	z	p	$\sigma_{\text{participant}}$
(Intercept)	-1.1370	0.0441	-25.81	< 0.0001	0.3775
Trial number	-0.0531	0.0067	-7.98	< 0.0001	0.0586
in_bnc=1	-0.1248	0.0060	-20.92	< 0.0001	0.0357
Word length	0.0246	0.0020	12.23	< 0.0001	0.0176
C-Precision	0.0544	0.0094	5.81	< 0.0001	0.0764
Yes-activation	-0.1517	0.0151	-10.05	< 0.0001	0.1302
Log Semantic density	-0.0450	0.0103	-4.35	< 0.0001	0.0870
response=W	-0.6712	0.0330	-20.36	< 0.0001	0.2817
Log Word frequency & in_bnc	-0.1100	0.0023	-47.74	< 0.0001	0.0200
Log L1Chat (response=N)	-0.1832	0.0108	-16.90	< 0.0001	0.0931
Log L1Chat (response=W)	0.0380	0.0107	3.55	0.0004	0.0928
Residual	0.3883				

(a) Words

	Est.	SE	z	p	$\sigma_{\text{participant}}$
(Intercept)	-0.8557	0.0423	-20.21	< 0.0001	0.3237
Trial number	-0.0628	0.0069	-9.08	< 0.0001	0.0610
Word length	0.0470	0.0018	26.37	< 0.0001	0.0154
has_neighbours_path=1	0.0718	0.0026	27.12	< 0.0001	0.0218
Yes-activation	0.1135	0.0162	6.99	< 0.0001	0.1406
response=W	-1.1454	0.0983	-11.65	< 0.0001	0.3793
Log Shortest Path & has_neighbours_path	0.0198	0.0012	16.90	< 0.0001	0.0091
Log L1Chat (response=N)	-0.3071	0.0129	-23.75	< 0.0001	0.1016
Log L1Chat (response=W)	0.0020	0.0294	0.07	0.9450	0.1306
Log Semantic density (response=N)	0.8793	0.0366	24.05	< 0.0001	0.2541
Log Semantic density (response=W)	0.0119	0.1090	0.11	0.9134	0.4063
Log L1Chat & Log Semantic density (response=N)	-0.2038	0.0096	-21.30	< 0.0001	0.0630
Log L1Chat & Log Semantic density (response=W)	-0.0133	0.0322	-0.41	0.6795	0.1212
Residual	0.3841				

(b) Nonwords

Table 6: Coefficient and random slope estimates in LMMs for words and nonwords respectively. Factors are treatment-coded.

model, with the exception of *Word length*, do not show strong effects of individual differences with respect to a subject's overall speed.

Moving on to the DLM-based predictors, shown in the second and third row of Figure 9a and Figure 9b. For ease of description, in the following we dichotomise the continuous scale of subject speed into slow subjects (left side of each plot) and fast subjects (right side of each plot). For the word model (Figure 9a) and fast subjects, we see stronger effects of most *lexical-DLM* measures (*Log L1Chat (response=W), C-Precision* and *Log Semantic density*), and only a weak effect of *Yes-activation* and *Log L1Chat (response=N)*. Slow subjects on the other hand show a stronger effect of *Yes-activation* and *Log L1Chat (response=N)*, and a weak or even reversed effect of the three *lexical-DLM* measures. For nonwords (see Figure 9b), the effects tend to be reversed. Fast subjects show a strong effect of *Yes-activation*, but only a weak effect of the *lexical-DLM* measures (in this case *Log Shortest Path*, *Log Semantic density (response=N)* and *Log L1Chat:Log Semantic density (response=N)*). Slow subjects show a strong effect of *lexical-DLM* measures and a weak one of *Yes-activation*. Only *Log L1Chat (response=N)* shows the same pattern regardless of lexicality, i.e. weak/strong influence for fast/slow subjects. An overview can be found in Table 7.

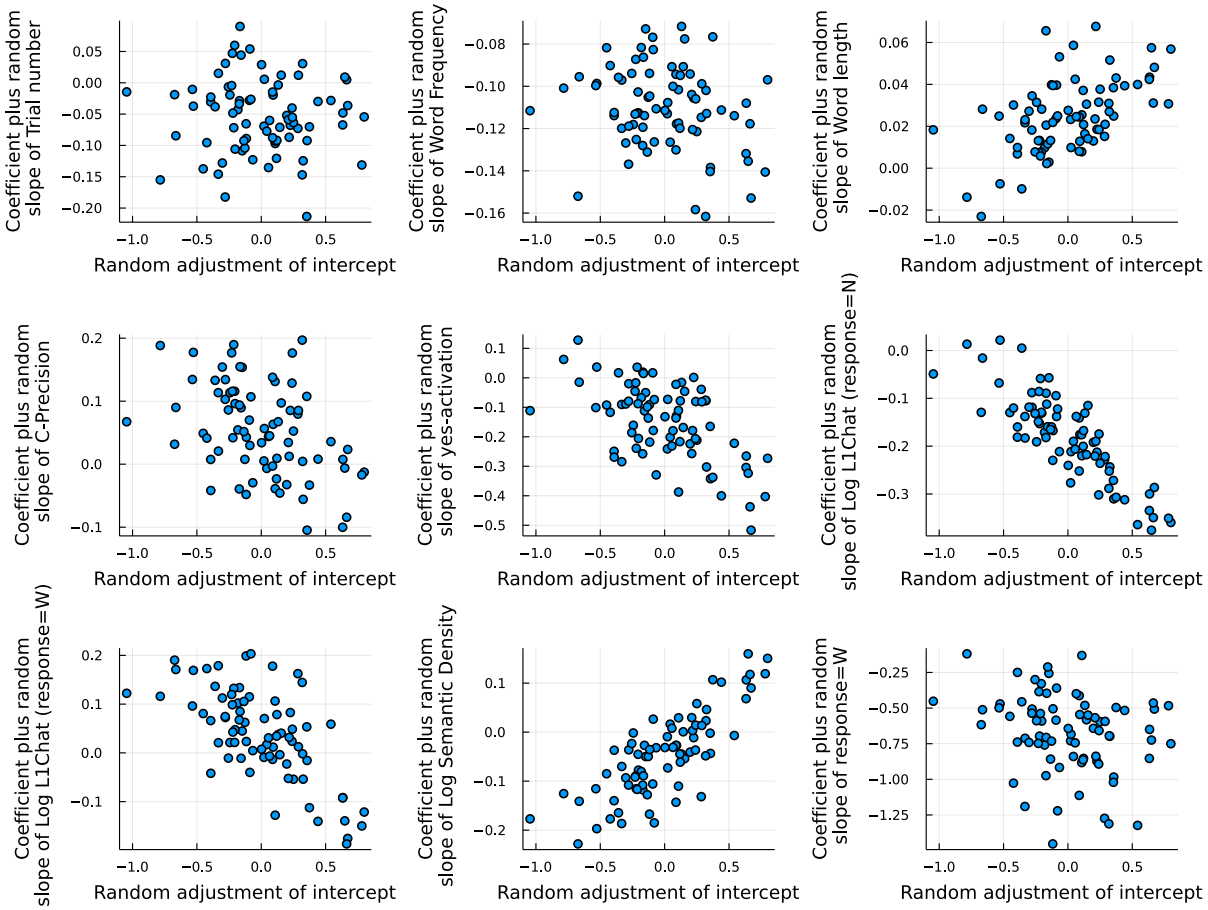
	Words		**Nonwords**	
	strong effect	weak/reversed effect	strong effect	weak/reversed effect
Slow subjects	*Word length*, *Yes-activation*, *Log L1Chat (response=N)*	*Log L1Chat (response=W)*, *Log Semantic density*, *C-Precision*	*Word length*, *Log Semantic density (response=N)*, *Log Shortest Path*, *Log L1Chat (response=N)*, *Log L1Chat:Log Semantic density (response=N)*	*Yes-activation*
Fast subjects	*Log L1Chat (response=W)*, *Log Semantic density*, *C-Precision*	*Word length*, *Yes-activation*, *Log L1Chat (response=N)*	*Yes-activation*	*Word length*, *Log Semantic density (response=N)*, *Log L1Chat (response=N)*, *Log Shortest Path*, *Log L1Chat:Log Semantic density (response=N)*

Table 7: Summary table of individual differences in predictor strengths, dichotomised into slow and fast subjects. For words, slow subjects show weak effects of predictors involving the DLM model, while fast subjects are strongly influenced by these predictors. The exact opposite pattern emerges for nonwords. The only exceptions to this pattern are *Word length* and *L1Chat (response=N)* which show a strong effect only for slow subjects across both words and nonwords. Non-significant predictors and predictors without clear correlation with the random effect of intercept are omitted.

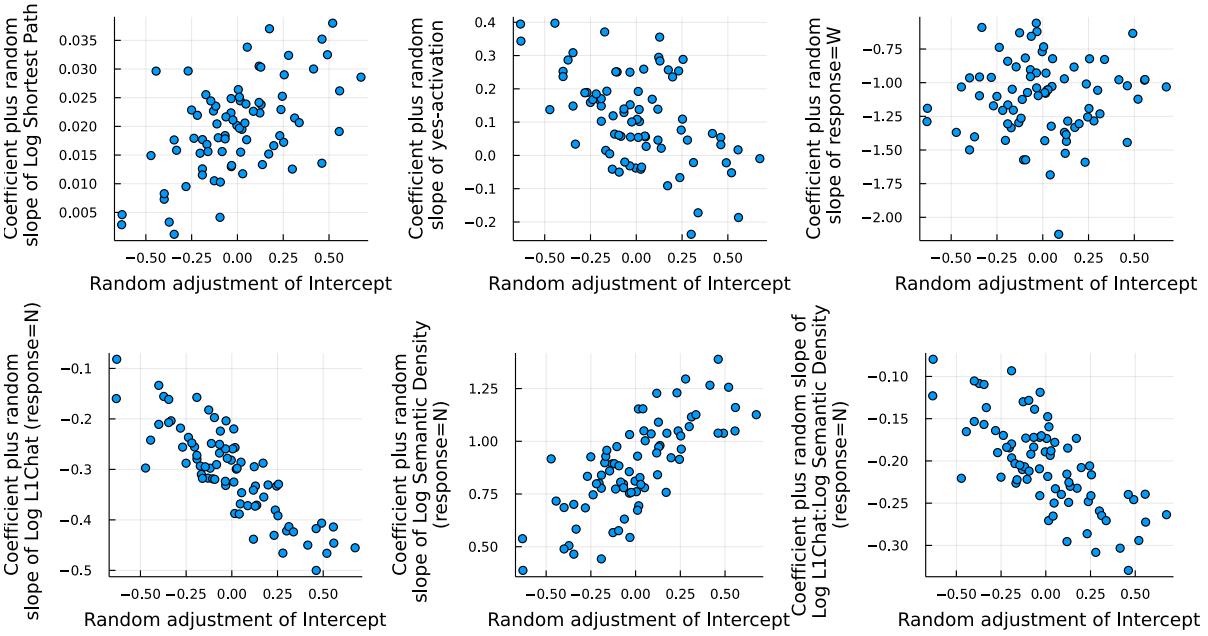
Apparently, slow subjects direct their attention to nonwords; this focus is reflected in large coefficients for measures that gauge deeper (incremental) lexical processing. Conversely, fast subjects appear to zoom in on words; as a consequence, their reaction times for words (rather than nonwords) show strong correlations with measures gauging deeper lexical processing. This interpretation is supported by the fact that the effect of *L1Chat (response=N)* is the same for both words and nonwords: slow subjects, attending to nonwords, focus strongly on *L1Chat (response=N)*, while fast subjects, attending to words, do not.

We note that the random slope adjustments of DLM-based predictors are correlated much more with each other than classical predictors (see Supplementary Materials). This could indicate that they all capture a similar dimension of individual difference, which is different from the one captured by, for instance, *Log Word frequency*.

We can therefore conclude that the DLM-based predictors provide a more precise window into investigating individual differences across speakers when it comes to the relationship between processing measures and overall speed than the non-incremental, classical measures (e.g. *word frequency*).



(a) Words



(b) Nonwords

Figure 9: Correlation between selection of participant-specific coefficients and random adjustment of intercept.

7 Discussion

The main question of the current study was whether the effects of within-experiment learning can be detected in the British Lexicon Project (BLP), a large-scale lexical decision experiment, using incremental learning. We investigated this question by simulating full lexical decision experiments for each subject in the BLP. We used a model of the mental lexicon called Discriminative Lexicon Model (DLM), which models comprehension and production with the help of simple mappings between form and meaning. It employs error-driven learning to compute its mappings. This allowed us to continuously update mappings between form and meaning throughout the simulation of each experiment. We then used learning-based predictors from the model to predict reaction times using Generalised Additive Models (GAMs), and found that for the majority of subjects predictors from a learning DLM are more successful at predicting reaction times than the same measures based on a non-learning DLM. We therefore conclude that trial-to-trial learning effects on reaction times in the BLP can be detected with error-driven learning.

Furthermore, we found that the DLM-based measures provided a substantially better model fit to lexical decision reaction times than models based on classical, non-incremental psycholinguistic measures (*Word frequency*, *Neighbourhood size*, *Word length*) for virtually all subjects. This adds to a range of previous studies demonstrating the ability of the DLM to predict behavioural data (Schmitz et al., 2021; Stein and Plag, 2021; Gahl and Baayen, 2022), contributing to Chuang and Baayen (2021)'s "external validation" of the DLM. For words, we found that participants responded faster if a word landed in dense semantic space and had high sublexical support for a word outcome. Both of these measures provide strong evidence in favour of "wordlikeness". Moreover, we found strong support for measures coming from an internal "feedback loop" (Chuang et al., 2020b) from meaning back to form: higher uncertainty (for word responses) and higher precision in this mapping were associated with longer reaction times. Our interpretation of the latter is that higher precision requires stronger suppression of the production system, which in turn slows down decisions.

The measures obtained for nonwords were highly consistent across subjects. We found that if a stimulus had many orthographic neighbours whose predicted semantic vectors were far apart in semantic space, reaction times were longer. If a stimulus received high sublexical support (evidence in favour of "wordlikeness"), reaction times were likewise longer. We also found an interaction between semantic density and lexical uncertainty in the feedback loop, especially for nonword responses: high semantic density and low lexical uncertainty were associated with longer reaction times. The results from the nonword models were therefore consistent with the results obtained for words: high sublexical support for word, low lexical uncertainty and high semantic density are all strong evidence in favour of a word response, and therefore slowed down nonword responses.

Unsurprisingly, these effects are not uniform across all subjects. First, language users have different exposure to and experience with language (Gardner et al., 1987; Keuleers et al., 2015; Ramscar, 2016; Hernandez et al., 2021). Second, cognitive differences may affect lexical processing (e.g. Kuperman and Van Dyke, 2011; Milin et al., 2017a; Fischer-Baum et al., 2018; Perfetti et al., 2005; Lõo et al., 2019). And third, the regression weight of lexical-distributional predictors can vary significantly between participants (example in Baayen, 2014). In the present study we found individual differences in our GAMs. We quantified this finding more formally using Linear Mixed Models, under the simplifying assumption that predictor effects are strictly linear. Slow and fast subjects differed in the precise effects of some of the predictors. For example, for slower subjects, the magnitude of the coefficient for word length was greater than for faster subjects. For words and slower subjects we found smaller absolute coefficients of measures of deeper, lexical processing (such as *Semantic density*) and bigger ones of high sublexical support for a word outcome (*Yes-activation*). For fast subjects on the other hand we observed smaller absolute coefficients of *Yes-activation*, and larger ones of deeper lexical processing such as *Se-*

mantic density (for fast subjects, high *Semantic density* of a stimulus elicits shorter reaction times) or lexical uncertainty (*L1Chat*). For nonwords the effects were reversed: slow subjects showed larger absolute coefficients of lexical processing measures, while fast subjects had larger ones of *Yes-activation*. Apparently, fast subjects focus their attention on words and therefore have big coefficients of deeper, lexical processing measures for words, whereas slow subjects focus their attention on nonwords.

Studying trial-to-trial effects with the DLM has various advantages: as mentioned above, it is able to model continuous learning by means of error-driven learning, which has been shown to be able to account for priming effects (Marsolek, 2008; Oppenheim et al., 2010). Priming effects have been modelled previously with the DLM, albeit without modelling trial-to-trial updates (Baayen and Smolka, 2020; Chuang et al., 2022b). Moreover, the DLM makes uses of (modality-specific) linear mappings for production and comprehension, and allows meaning and form to inform each other. Meanings are represented by means of high-dimensional vectors which are able to capture fine-grained meaning similarities, and have been used to model lexical processing in previous work (e.g. Mandera et al., 2017; Westbury and Wurm, 2022). This means that the DLM is able to move beyond the hand-crafted localist representation of semantics employed in NDL (Baayen et al., 2011; Norris, 2013). Measures of the interaction between form and meaning have been found to be predictive for lexical decision reaction times in previous studies, for example with the Orthography-to-Semantics Consistency measure by Marelli et al. (2015). Such interacting effects between form and meaning arise naturally in the DLM.

For the present simulation of lexical decision with the DLM, a range of modelling decisions had to be made. For example, the modelling of nonword trials posed challenges. Chuang et al. (2020b) showed that the semantics obtained for nonwords by projecting their orthographical representations into semantic space are predictive for reaction times and acoustic durations. However, they did not model learning in individual (nonword) trials as required in our study. Error-driven learning requires a target for each update, which in turn means that semantic targets are required for nonword trials. We modeled nonword semantics as a continuously updated and ever changing location in semantic space, different across subjects, and within subjects updated from nonword trial to nonword trial.

A further challenge included how to precisely measure processing within the DLM — similar to classical measures such as frequency, word length or age of acquisition, many different measures from the DLM can be considered as predictors for reaction times. Because of such issues, we developed our simulations, measures and statistical models using the data of the first two subjects in the BLP only, and tested them with the remaining subjects. This solution was computationally lean, but presumably led to suboptimal prediction accuracy because of the immense variability across subjects, depending on their individual exposure to language (Gardner et al., 1987; Ramscar, 2016; Hernandez et al., 2021) as well as their irregular within-experiment learning trajectories. A future study might train on data from all subjects, and test on held-out data from all subjects.

More challenges remain for future research. Our present mappings are initialised as the so-called "endstate of learning", meaning that all words in the lexicon are learned equally well. This is of course not realistic — due to different amounts of exposure, some words are learned much better than other words. While the DLM is in theory able to account for this by incrementally initialising the mappings, this proved to be computationally not feasible in the present study. Therefore, we are currently developing techniques for approximating frequency-informed initial mappings. Another challenge is the development of a decision mechanism. The goal of the present study was to examine trial-to-trial learning in lexical processing, so we restricted ourselves to modelling reaction times by predicting them directly from the measures gauged from the DLM. This follows theories from previous work such as Redgrave et al. (1999) and Gurney et al. (2001) that decisions are made by distinct cognitive control processes and hence are not a component of processing in the mental lexicon. Architectures such as ACT-R (Ander-

son and Lebiere, 1998) or, at a lower level, PRIMS (Taatgen, 2013), might be able to combine information from lexical processing and learning with cognitive control processes.

In conclusion, even though the equations underlying LDL are extremely simple, they are nonetheless able to capture fine-grained trial-to-trial learning effects in large-scale studies of lexical decision such as the BLP. While previous work has shown the predictive power of error-driven learning by means of the Rescorla-Wagner rule in many areas of language research (e.g. Ramscar et al., 2013; Nixon and Tomaschek, 2021; Ellis, 2006a,b) as well as trial-to-trial learning (Chuang and Baayen, 2021; Lentz et al., 2021; Tomaschek et al., 2022), the present work extends these findings by showing that error-driven learning implemented with the Widrow-Hoff learning rule can also capture trial-to-trial learning effects in a much richer and more realistic model of lexical processing incorporating state-of-the-art representations of semantics.

Acknowledgments

This work has been supported by the European Research Council, project WIDE-742545. The authors thank Fabian Tomaschek and Tino Sering for comments on earlier versions of this manuscript and participants of the Groningen Spring School on Cognitive Modeling for their helpful feedback.

References

Aguasvivas, J. A., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., and Duñabeitia, J. A. (2018). Spalex: A spanish lexical decision database from a massive online data collection. *Frontiers in psychology*, 9:2156.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer.

Anderson, J. and Lebiere, C. (1998). *The Atomic Components of Thought (1st ed.)*. Psychology Press.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2):234.

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., and Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PloS one*, 12(4):e0174623.

Arnon, I. and Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, 122(3):292–305.

Baayen, H., Vasishth, S., Kliegl, R., and Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 94:206–234.

Baayen, R. H. (2005). Data mining at the intersection of psychology and linguistics. In *Twenty-first century psycholinguistics: Four cornerstones*, pages 69–84. Routledge.

Baayen, R. H. (2014). Multivariate statistics. *Research methods in linguistics*, page 337.

Baayen, R. H., Chuang, Y.-Y., and Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon*, 13(2):230–268.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity*, 2019.

Baayen, R. H., Fasiolo, M., Wood, S., and Chuang, Y.-Y. (2022). A note on the modeling of the effects of experimental time in psycholinguistic experiments. *The Mental Lexicon*.

Baayen, R. H., Feldman, L. B., and Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 55(2):290–313.

Baayen, R. H., Hendrix, P., and Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and speech*, 56(3):329–347.

Baayen, R. H., Milin, P., and Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, 30(11):1174–1220.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review*, 118(3):438.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX lexical database [cd rom]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Baayen, R. H. and Smolka, E. (2020). Modeling morphological priming in german with naive discriminative learning. *Frontiers in Communication*, 5:17.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., and Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of experimental psychology: General*, 133(2):283.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The english lexicon project. *Behavior research methods*, 39(3):445–459.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247.

Bates, D., Alday, P., Kleinschmidt, D., Calderón, J. B. S., Zhan, L., Noack, A., Arslan, A., Bouchet-Valat, M., Kelman, T., Baldassari, A., Ehinger, B., Karrasch, D., Saba, E., Quinn, J., Hatherly, M., Piibeleht, M., Mogensen, P. K., Babayan, S., and Gagnon, Y. L. (2021). Juliastats/mixedmodels v4.3.0.

Bennett, D., Murawski, C., and Bode, S. (2015). Single-trial event-related potential correlates of belief updating. *ENeuro*, 2(5).

Bowers, J. S., Davis, C. J., and Hanley, D. A. (2005). Automatic semantic activation of embedded words: Is there a "hat" in "that"? *Journal of Memory and Language*, 52(1):131–143.

Bröker, F. and Ramscar, M. (2020). Representing absence of evidence: why algorithms and representations matter in models of language and cognition. *Language, Cognition and Neuroscience*, pages 1–24.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4):977–990.

Brysbaert, M., Stevens, M., Mandera, P., and Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the dutch lexicon project 2. *Journal of Experimental Psychology: Human Perception and Performance*, 42(3):441.

Buchanan, L., Westbury, C., and Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, 8(3):531–544.

Cassani, G., Chuang, Y.-Y., and Baayen, R. H. (2019). On the semantics of non-words and their lexical category. *Journal of Experimental Psychology: Learning, Memory, Cognition*, 46(4):621–637.

Chang, F., Dell, G. S., and Bock, K. (2006). Becoming syntactic. *Psychological review*, 113(2):234.

Chang, Y.-N., Ralph, M. L., Furber, S., and Welbourne, S. (2013). Modelling graded semantic effects in lexical decision. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Chuang, Y.-Y. and Baayen, R. H. (2021). Discriminative learning and the lexicon: Ndl and ldl. In *Oxford Research Encyclopedia of Linguistics*.

Chuang, Y.-Y., Kang, M., Luo, X., and Baayen, R. H. (2022a). Vector space morphology with linear discriminative learning. In Crepaldi, D., editor, *Linguistic morphology in the mind and brain*.

Chuang, Y. Y., Kang, M., Luo, X. F., and Baayen, R. H. (2022b). Vector space morphology with linear discriminative learning. In Crepaldi, D., editor, *Linguistic morphology in the mind and brain*. Routledge.

Chuang, Y.-Y., Lõo, K., Blevins, J. P., and Baayen, R. H. (2020a). Estonian case inflection made simple a case study in word and paradigm morphology with. In Körtvélyessy, L. and Štekauer, P., editors, *Complex Words: Advances in Morphology*, chapter 7, pages 119–14. Cambridge University Press.

Chuang, Y.-Y., Vollmer, M.-l., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., and Baayen, R. H. (2020b). The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behaviour Research Methods*.

Chumbley, J. I. and Balota, D. A. (1984). A word's meaning affects the decision in lexical decision. *Memory & Cognition*, 12(6):590–606.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee.

Danks, D. (2003). Equilibria of the rescorla–wagner model. *Journal of Mathematical Psychology*, 47(2):109–121.

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological review*, 117(3):713.

Dell, G. S. and Caramazza, A. (2008). Introduction to special issue on computational modelling in cognitive neuropsychology. *Cognitive Neuropsychology*, 25(2):131–135.

Diedrichsen, J., White, O., Newman, D., and Lally, N. (2010). Use-dependent and error-based learning of motor behaviors. *Journal of Neuroscience*, 30(15):5159–5166.

Dijkstra, T. and Van Heuven, W. J. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and cognition*, 5(3):175–197.

Ellis, N. C. (2006a). Language acquisition as rational contingency learning. *Applied linguistics*, 27(1):1–24.

Ellis, N. C. (2006b). Selective attention and transfer phenomena in l2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied linguistics*, 27(2):164–194.

Ellis, N. C. and Sagarra, N. (2010). The bounds of adult language acquisition: Blocking and learned attention. *Studies in Second Language Acquisition*, 32(4):553–580.

Engbert, R., Longtin, A., and Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision research*, 42(5):621–636.

Fischer-Baum, S., Kook, J. H., Lee, Y., Ramos-Nuñez, A., and Vannucci, M. (2018). Individual differences in the neural and cognitive mechanisms of single word reading. *Frontiers in human neuroscience*, 12:271.

Forster, K. I. and Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of experimental psychology: Learning, Memory, and Cognition*, 10(4):680.

Friedman, L. and Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, 59(2):127–136.

Gahl, S. and Baayen, R. H. (2022). Time and thyme again: Connecting spoken word duration to models of the mental lexicon. *https://osf.io/kxpaj/*.

Gardner, M. K., Rothkopf, E. Z., Lapan, R., and Lafferty, T. (1987). The word frequency effect in lexical decision: Finding a frequency-based component. *Memory and Cognition*, 15:24–28.

Grainger, J. and Jacobs, A. M. (1996). Orthographic processing in visual word recognition: a multiple read-out model. *Psychological review*, 103(3):518.

Günther, F., Rinaldi, L., and Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, 14(6):1006–1033.

Gurney, K., Prescott, T., and Redgrave, P. (2001). A computational model of action selection in the basal ganglia. *Biol. Cybern.*, 84:401–423.

Harm, M. W. and Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological review*, 111(3):662.

Harris, Z. S. (1954). Distributional Structure. *WORD*, 10(2-3).

Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386.

Heitmeier, M. and Baayen, R. H. (2020). Simulating phonological and semantic impairment of English tense inflection with linear discriminative learning. *The Mental Lexicon*, 15(3):385–421.

Heitmeier, M., Chuang, Y.-Y., and Baayen, R. H. (2021). Modeling morphology with linear discriminative learning: considerations and design choices. *Frontiers in psychology*, page 4929.

Hendrix, P. and Sun, C. C. (2021). A word or two about nonwords: Frequency, semantic neighborhood density, and orthography-to-semantics consistency effects for nonwords in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(1):157.

Hernandez, A. E., Ronderos, J., Bodet, J. P., Claussenius-Kalman, H., Nguyen, M. V., and Bunta, F. (2021). German in childhood and latin in adolescence: On the bidialectal nature of lexical access in english. *Humanities and Social Sciences Communications*, 8(1):1–12.

Hopman, E., Thompson, B., Austerweil, J., and Lupyan, G. (2018). Predictors of l2 word learning accuracy: A big data investigation. In *the 40th Annual Conference of the Cognitive Science Society (CogSci 2018)*, pages 513–518. Cognitive Science Society.

Hoppe, D. B., Hendriks, P., Ramscar, M., and van Rij, J. (2022). An exploration of error-driven learning in simple two-layer networks from a discriminative learning perspective. *Behavior Research Methods*, pages 1–31.

Hoppe, D. B., van Rij, J., Hendriks, P., and Ramscar, M. (2020). Order matters! influences of linear order on linguistic category learning. *Cognitive Science*, 44(11):e12910.

Jacobs, A. M. and Grainger, J. (1994). Models of visual word recognition: sampling the state of the art. *Journal of Experimental Psychology: Human perception and performance*, 20(6):1311.

Kell, C. A., Darquea, M., Behrens, M., Cordani, L., Keller, C., and Fuchs, S. (2017). Phonetic detail and lateralization of reading-related inner speech and of auditory and somatosensory feedback processing during overt reading. *Human brain mapping*, 38(1):493–508.

Keuleers, E. and Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior research methods*, 42(3):627–633.

Keuleers, E., Lacey, P., Rastle, K., and Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior research methods*, 44(1):287–304.

Keuleers, E., Stevens, M., Mandera, P., and Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology*, (8):1665–1692.

Kuperman, V. and Van Dyke, J. (2011). Individual differences in visual comprehension of morphological complexity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Lõo, K., Toth, A., Karaca, F., and Järvikivi, J. (2019). Effects of affective ratings and individual differences in English morphological processing. In *CogSci*, pages 2179–2185.

Lentz, T., Nixon, J. S., and van Rij, J. (2021). Temporal response modelling uncovers electrophysiological correlates of trial-by-trial error-driven learning. *PsyArXiv: 10.31234/osf.io/dg5mw*.

Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1):1–36.

Linke, M., Bröker, F., Ramscar, M., and Baayen, H. (2017). Are baboons learning" orthographic" representations? probably not. *PloS one*, 12(8):e0183876.

Luo, X. (2021). JudiLing: An implementation for Linear Discriminative Learning in JudiLing (unpublished Master's thesis). *https://github.com/MegamindHenry/JudiLing.jl/blob/master/thesis/thesis_JudiLing__An_implementation_for_Discriminative_Learning_in_Julia.pdf*.

Mandera, P., Keuleers, E., and Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78.

Marelli, M. and Amenta, S. (2018). A database of orthography-semantics consistency (osc) estimates for 15,017 english words. *Behavior research methods*, 50(4):1482–1495.

Marelli, M., Amenta, S., and Crepaldi, D. (2015). Semantic transparency in free stems: The effect of orthography-semantics consistency on word recognition. *Quarterly Journal of Experimental Psychology*, 68(8):1571–1583.

Marsolek, C. J. (2008). What antipriming reveals about priming. *Trends in Cognitive Sciences*, 12(5):176–181.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, 1(1):11–38.

McClelland, J. L. and Rumelhart, D. E. (1989). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. MIT press.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Milin, P., Divjak, D., and Baayen, R. H. (2017a). A learning perspective on individual differences in skilled reading: Exploring and exploiting orthographic and semantic discrimination cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(11):1730.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., and Baayen, R. H. (2017b). Discrimination in lexical decision. *PloS one*, 12(2):e0171935.

Nassar, M. R., Wilson, R. C., Heasly, B., and Gold, J. I. (2010). An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, 30(37):12366–12378.

New, B., Ferrand, L., Pallier, C., and Brysbaert, M. (2006). Re-examining word length effects in visual word recognition: New evidence from the english lexicon project. *Psychonomic bulletin & review*, 13:45–52.

Nieder, J., Chuang, Y.-Y., van de Vijver, R., and Baayen, R. (2021). Comprehension, production and processing of maltese plurals in the discriminative lexicon.

Nixon, J. S. and Tomaschek, F. (2021). Prediction and error in early infant speech learning: A speech acquisition model. *Cognition*, 212:104697.

Norris, D. (2006). The bayesian reader: explaining word recognition as an optimal bayesian decision process. *Psychological review*, 113(2):327.

Norris, D. (2013). Models of visual word recognition. *Trends in cognitive sciences*, 17(10):517–524.

Oppenheim, G. M., Dell, G. S., and Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, 114(2):227–252.

O'Reilly, R. C., Russin, J. L., Zolfaghar, M., and Rohrlich, J. (2021). Deep predictive learning in neocortex and pulvinar. *Journal of Cognitive Neuroscience*, 33(6):1158–1196.

O'Reilly, R. and Rohrlich, J. (2018). Deep predictive learning in vision. In *2018 Conference on Cognitive Computational Neuroscience*, pages 2018–1242.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Perfetti, C. A., Wlotko, E. W., and Hart, L. A. (2005). Word learning and individual differences in word learning reflected in event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6):1281.

Pferschy, U. and Staněk, R. (2017). Generating subtour elimination constraints for the tsp from pure integer solutions. *Central European journal of operations research*, 25(1):231–260.

Pham, H. and Baayen, H. (2015). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition and Neuroscience*, 30(9):1077–1095.

Pritchard, S. C., Coltheart, M., Marinus, E., and Castles, A. (2016). Modelling the implicit learning of phonological decoding from training on whole-word spellings and pronunciations. *Scientific studies of reading*, 20(1):49–63.

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, 103(20):7865–7870.

Ramscar, M. (2016). Learning and the replicability of priming effects. *Current Opinion in Psychology*, 12:80–84.

Ramscar, M., Dye, M., and McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of "mouses" in adult speech. *Language*, pages 760–793.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., and Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in cognitive science*, 6(1):5–42.

Ramscar, M., Sun, C. C., Hendrix, P., and Baayen, H. (2017). The mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the "cost" of learning, not cognitive decline. *Psychological science*, 28(8):1171–1179.

Ramscar, M. and Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive science*, 31(6):927–960.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., and Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive science*, 34(6):909–957.

Ratcliff, R., Gomez, P., and McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological review*, 111(1):159.

Redgrave, P., Prescott, T., and Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89:1009–1023.

Rescorla, R. A. and Wagner, A. R. (1972). *Classical conditioning II: current research and theory*, chapter A theory of Pavlovian conitioning: variations in the effectiveness of reinforcement and nonreinforcement, pages 64–99. Appleton-Century-Crofts, New York.

Rodd, J. M. (2004). When do leotards get their spots? semantic activation of lexical neighbors in visual word recognition. *Psychonomic Bulletin & Review*, 11(3):434–439.

Roediger, H. L. (1993). Implicit memory in normal subjects. *Handbook of neuropsychology*, 8:63–131.

Rubenstein, H., Garfield, L., and Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of verbal learning and verbal behavior*, 9(5):487–494.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning by error backpropagation. In *Parallel Distributed Processing*, volume 1. MIT press.

Rumelhart, D. E. and McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Ii. the contextual enhancement effect and some tests and extensions of the model. *Psychological review*, 89(1):60.

Scarborough, D. L., Cortese, C., and Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human perception and performance*, 3(1):1.

Schmitz, D., Plag, I., Baer-Henney, D., and Stein, S. D. (2021). Durational differences of word-final/s/emerge from the lexicon: Modelling morpho-phonetic effects in pseudowords with linear discriminative learning. *Frontiers in Psychology*, 12:2983.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80:1–27.

Seidenberg, M. and McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4):523.

Shafaei-Bajestan, E., Moradipour-Tari, M., Uhrig, P., and Baayen, R. H. (2021). Ldl-auris: a computational model, grounded in error-driven learning, for the comprehension of single spoken words. *Language, Cognition and Neuroscience*, pages 1–28.

Shahmohammadi, H., Lensch, H. P., and Baayen, H. (2021). Learning zero-shot multifaceted visually grounded word embeddings via multi-task training. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 158–170.

Skipper, J. I., Devlin, J. T., and Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and language*, 164:77–105.

Stein, S. D. and Plag, I. (2021). Morpho-phonetic effects in speech production: Modeling the acoustic duration of english derived words with linear discriminative learning. *Frontiers in Psychology*, 12.

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological review*, 120(3):439.

Tomaschek, F., Hendrix, P., and Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, 71:249–267.

Tomaschek, F., Ramscar, M., and Nixon, J. (2022). The keys to the future? An examination of associative versus discriminative accounts of Serial Pattern Learning. *Submitted to Cognitive Science*.

Trimmer, P. C., McNamara, J. M., Houston, A. I., and Marshall, J. A. R. (2012). Does natural selection favour the Rescorla-Wagner rule? *Journal of Theoretical Biology*, 302:39–52.

Van Rijn, H. and Anderson, J. R. (2003). Modeling lexical decision as ordinary retrieval. *Department of Psychology*.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., and McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of memory and language*, 58(1):140–159.

Westbury, C., Keith, J., Briesemeister, B. B., Hofmann, M. J., and Jacobs, A. M. (2014). Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. *The Quarterly Journal of Experimental Psychology*.

Westbury, C. and Wurm, L. H. (2022). Is it you you're looking for? Personal relevance as a principal component of semantics. *The Mental Lexicon*, 17(1):1–33.

Widrow, B. and Hoff, M. (1960). Adaptive switching circuits. *1960 WESCON Convention Record Part IV*.

Wilson, R. C. and Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *Elife*, 8:e49547.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.

Yap, M. J., Sibley, D. E., Balota, D. A., Ratcliff, R., and Rueckl, J. (2015). Responding to nonwords in the lexical decision task: Insights from the english lexicon project. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(3):597.