Frequency effects in Linear Discriminative Learning

Maria Heitmeier[1,2], Yu-Ying Chuang[1], Seth D. Axen[2], and R. Harald Baayen[1,2]

[1]Quantitative Linguistics, University of Tübingen
[2]Cluster of Excellence Machine Learning: New Perspectives for Science, University of Tübingen

arXiv:2306.11044v1 [cs.CL] 19 Jun 2023

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 This paper has not been peer reviewed.

Word frequency is a strong predictor in most lexical processing tasks (Brysbaert et al., 2011). Thus, any model of word recognition needs to account for how word frequency effects arise. The Discriminative Lexicon Model (DLM; Baayen et al., 2018a, 2019) models lexical processing with mappings between words' forms and their meanings. Comprehension and production are modelled via linear mappings between the two domains. So far, the mappings within the model can either be obtained incrementally via error-driven learning, a computationally expensive process able to capture frequency effects, or in an efficient, but frequency-agnostic closed-form solution modelling the theoretical endstate of learning (EL) where all words are learned optimally. In the present study we show how an efficient, yet frequency-informed mapping between form and meaning can be obtained (Frequency-informed learning; FIL). We find that FIL well approximates an incremental solution while being computationally much cheaper. FIL shows a relatively low type- and high token-accuracy, demonstrating that the model is able to process most word tokens encountered by speakers in daily life correctly. We use FIL to model reaction times in the Dutch Lexicon Project (Keuleers et al., 2010) by means of a Gaussian Location Scale Model and find that FIL predicts well the S-shaped relationship between frequency and the mean of reaction times but underestimates the variance of reaction times for low frequency words. FIL is also better able to account for priming effects in an auditory lexical decision task in Mandarin Chinese (Lee, 2007), compared to EL. Finally, we used ordered data from CHILDES (Brown, 1973; Demuth et al., 2006) to compare mappings obtained with FIL and incremental learning. We show that the mappings are highly correlated, but that with FIL some nuances based on word ordering effects are lost. Our results show how frequency effects in a learning model can be simulated efficiently by means of a closed-form solution, and raise questions about how to best account for low-frequency words in cognitive models.

Keywords: linear discriminative learning, word frequency, incremental learning, weighted regression, lexical decision, mental lexicon, distributional semantics

1 Introduction

Word frequency effects are ubiquitous in psycholinguistic research. In fact, word frequency (i.e. the number of times a word occurs in some corpus) is one of the most important predictors in a range of psycholinguistic experiments (Brysbaert et al., 2011). In the lexical decision task, where participants are asked to decide whether a presented letter string is a word or not, frequency explains by far the most variance in reaction times, compared to other measures such as neighbourhood

density or word length (e.g. Baayen, 2010; Brysbaert et al., 2011): higher frequency words elicit faster reactions (e.g. Rubenstein et al., 1970; Balota et al., 2004). In word naming, another popular experiment in psycholinguistics where participants have to read aloud presented words, word frequency is less important, but still has a reliable effect: higher frequency words are named faster (e.g. Balota et al., 2004). Even though the effect of frequency has long been known and studied, to this day new studies are published confirming the effect in ever larger datasets across different languages (e.g. Balota et al., 2007; Keuleers et al., 2012, 2010; Aguasvivas et al., 2018; Brysbaert et al., 2016; Ferrand et al., 2010) and proposing new frequency counts, explaining for example individual differences when it comes to the frequency effect (e.g. Brysbaert et al., 2018; Kuperman and Van Dyke, 2013).

Unsurprisingly, a central challenge for models of word recognition has therefore always been to explain word frequency effects. This challenge has been met in various different ways by influential models of word recognition. One of the earliest ideas proposed that words are stored in list-like structures ordered by frequency, such that the most frequent words are found earlier than lower frequency words (e.g. Rubenstein et al., 1971). This idea was developed into the "Search Model of Lexical Access" by Forster (1976, 1979). In this model there are "peripheral access files" in which words are stored according to their frequency of occurrence. These files hold "access keys" to where information about each word is stored in a main file. Words in the access files are grouped into bins based on form characteristics. Thus, the model aims to explain both word frequency effects (words are ordered according to their frequency in the peripheral access files) as well as neighbourhood effects (words with similar form are stored in the same bin). Later iterations of the model also suggested a hybrid model between serial and parallel search, where each bin is searched serially, but all bins are searched in parallel (Forster, 1994).

The Logogen model (Morton, 1969, 1979a,b) fully doubled down on the idea of a parallel search. A logogen can be seen as a detector for a set of input features. Every time one of its associated input features is encountered, the logogen's counter is increased. If the counter surpasses a threshold, it elicits a response. Each word/morpheme in a speaker's lexicon is assumed to correspond to a logogen. Additionally, there are logogens for lower level visual and auditory input features such as letters or phonemes whose outputs in turn serve as inputs to the word logogens. The Logogen model accounts for frequency effects by assuming that logogens corresponding to words with higher frequency have a lower response threshold than those corresponding to lower frequency words. After a logogen elicits a response, the threshold is lowered, and it takes a long time for the threshold to increase again. This explains trial-to-trial effects because just activated words will be activated faster in subsequent trials, but it also explains the long-term effect of word frequency because words occurring regularly will always have a lower threshold for eliciting a response (Morton, 1969).

The interactive activation model (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982) is in many ways a successor of the Logogen model. It proposes three levels of representations: one for letter features, one for letters and one for words. There are excitatory and inhibitory connections from letter features to letters and from letters to words. Additionally, there are excitatory and inhibitory connections from words to letters. Finally, there are both excitatory and inhibitory connections within each representational level. The interactive activation model was originally proposed to explain the word superiority effect, i.e. the finding that letters are identified faster within a word than within a random letter string (e.g. Reicher, 1969). However, the model of course also proposes an account of frequency effects: the resting activation level of the word units are set depending on word frequency, such that high frequency words have a higher resting activation than low frequency words. In this way, Jacobs and Grainger (1994) were able to show that the interactive activation model shows the same effect of frequency on reaction times in various lexical decision experiments as human participants.

The interactive activation model was followed by the triangle model (Seidenberg and McClelland, 1989; Harm and Seidenberg, 2004), which consists of distributed representations for orthography, phonology and semantics, each connected via a set of hidden units. In contrary to the interactive activation model, the weights between layers in this network are not set by the modeller but learned using backpropagation (Rumelhart et al., 1986). Thus, the error between the model's prediction

and the actual target is reduced each time an input-target pair is encountered. For example, to model reading, the triangle model predicts a word's phonology from its orthography. Then, the error between the predicted phonology and the correct phonology of the word is computed, and the model's weights are updated such that the next time the same phonology is to be produced from the same orthography, it will be more accurate. This means that words that are presented to the model more frequently will over time produce the lowest prediction error and are thus recognised faster and more accurately. Therefore, word frequency effects arise not as a consequence of manually changing resting activation levels but from the weights within the network changing according to the input distribution.

A final model of word recognition is the Bayesian Reader Model (Norris, 2006). This model not only accounts for the frequency effect but also aims to explain **why** frequency effects should arise in the first place. The model is a simple Bayesian model that integrates a word's prior probability (for which Norris, 2006, uses its frequency) with the incoming evidence. Thus, high frequency words are recognised faster than low frequency ones. According to Norris (2006, 2013) this constitutes an "ideal observer" model, solving the task at hand as optimally as possible. This explains not only why frequency effects should arise in the first place but also why they play out differently in different experiments.

To summarise, these five models offer three broad explanations of frequency effects. Serial search models explain them in terms of list ordering effects; the Logogen, interactive activation and triangle model propose network models where frequency is reflected in units' thresholds/activation levels or in connection weights; and finally, the Bayesian Reader proposes that word frequencies provide lexical priors that contribute to an optimal decision process in word recognition.

Interestingly, reaction times for example in lexical decision are best predicted not by raw word frequencies but by log- or rank-transformed frequencies. Again, the various models account for this in different ways. Since the serial search model incorporates word lists, it directly predicts a rank transformation of word frequencies (see also Murray and Forster, 2004). In the interactive activation model, the resting activation levels of the word units are set according to the words' log frequency (McClelland and Rumelhart, 1989, Chapter 7). In the triangle model, training items are sampled such that they have a probability proportional to $\log(\text{frequency}+2)$. Note that Seidenberg and McClelland (1989) do this for practical rather than principled reasons, as training their model on the full token frequencies was too computationally expensive. Finally, the Bayesian Reader utilises raw word frequencies.

Comparison of these models highlights a few key questions about how to model the word frequency effect: first, how and why does the frequency effect arise in the first place? Does it arise naturally as a consequence of the input data? And what mechanism does the model provide for how the frequency differences are acquired? Secondly, how does the model keep track of word frequencies? Are there "counters" for each individual word (see also Baayen, 2010)? And finally, how does the model account for the fact that reaction times are best predicted by log- or rank-transformed frequencies rather than raw frequency counts?

We now turn to a newer model of word comprehension and production, the Discriminative Lexicon Model (DLM; Baayen et al., 2018a, 2019). This model provides a perspective on the mental lexicon in which mappings between numeric representations for form and meaning play a central role. This model conceptualizes comprehension as involving mappings from high-dimensional modality-specific form vectors to high-dimensional representations of meaning. The initial stage of speech production is modeled as involving a mapping in the opposite direction, starting with a high-dimensional semantic vector (known as embeddings in computational linguistics) and targeting a vector specifying which phone combinations drive articulation. The DLM has been successful in modelling a range of different morphological systems (e.g. Chuang et al., 2020, 2022; Denistia and Baayen, 2021; Nieder et al., 2021; Heitmeier et al., 2021) as well as behavioural data such as acoustic durations (Schmitz et al., 2021; Stein and Plag, 2021; Chuang et al., 2022), (primed) lexical decision reaction times (Gahl and Baayen, 2023; Heitmeier et al., 2022) and data from patients with aphasia (Heitmeier and Baayen, 2020).

The DLM's mappings between form and meaning are implemented by means of matrices. This

general matrix-based approach is referred to as linear discriminative learning (LDL). LDL can be implemented in two ways: by means of the matrix algebra underlying multivariate multiple regression (henceforth the 'endstate learning', EL), or by means of incremental regression using the error-driven learning rule of Widrow and Hoff (1960) (henceforth WHL). EL is computationally efficient, WHL is computationally demanding. Conversely, WHL is sensitive to the frequencies with which words are presented for learning, whereas EL is fully type-based (i.e. words' token frequencies do not play a role).

Thus, the DLM proposes that frequency effects arise due to the distribution of the input data: higher frequency words occur more often in the input data, and therefore, the prediction error will be smallest for high frequency words (see Chuang et al., 2021; Heitmeier et al., 2021, for studies utilising WHL to obtain frequency-informed mapping matrices). Word frequencies are not stored explicitly; rather, they have effects on the weights in the mappings. This account is similar to how frequency effects arise in the triangle model. Similar to the triangle model, the DLM also suffers from computational issues: training on the full frequency distribution with WHL is computationally very demanding. However, the DLM has an advantage over the triangle model: EL only involves a single linear mapping for which a closed-form solution exists. In the present study we introduce a new method which takes frequency of use into account but is computationally efficient by making use of a closed-form solution ("Frequency-informed learning", FIL). FIL can be used instead of the already established WHL and EL to compute mapping matrices in the DLM.

In the following we compare the three different methods of estimating mapping matrices in the DLM. We show how the model is able to account for frequency effects using WHL and FIL. We demonstrate that FIL is equivalent to training the model incrementally on full token frequencies and is superior to utilising log-transformed frequencies. We show how the DLM is able to model reaction times linearly without the need of log- or rank-transformations. Finally, we investigate what role the order in which words are learned plays in word recognition.

This study is structured as follows: we first lay out the basic concepts underlying linear discriminative learning. Next, we discuss EL, and subsequently, WHL. Against this background, we proceed with proposing a new method computing non-incremental, yet frequency-informed learning (FIL). We then present three case studies, one using FIL to model visual lexical decision latencies from the Dutch Lexicon Project (Keuleers et al., 2010), one where we use FIL to model spoken word recognition in Mandarin and a third where we compare WHL and FIL in modelling first word learning with data from CHILDES (Brown, 1973; Demuth et al., 2006). A discussion section brings this study to a close.

2 Linear Discriminative Learning: Basic concepts and notation

In the DLM, word forms are represented as binary vectors that code the presence and absence of overlapping n-grams in the word form (e.g., `#a`, `aa`, `ap`, `p#` for the Dutch word *aap*, '*monkey*'). Form vectors are stored as row vectors in a 'cue matrix' $\mathbf{C}$. For an overview of how form vectors can be constructed, see Heitmeier et al. (2021), and for form vectors derived from audio signals, see Shafaei-Bajestan et al. (2021). Semantics are represented as real-valued vectors, following distributional semantics (Landauer et al., 1998), which are stored as row vectors in a semantic matrix $\mathbf{S}$.

To model comprehension, a mapping matrix $\mathbf{F}$ transforms the form vectors in $\mathbf{C}$ into the semantic vectors in $\mathbf{S}$. Conversely, a production matrix $\mathbf{G}$ maps meanings onto forms. The matrices $\mathbf{F}$ and $\mathbf{G}$ are estimated by solving

$$\begin{aligned} \mathbf{CF} &= \mathbf{S} \\ \mathbf{SG} &= \mathbf{C}. \end{aligned}$$

Given $\mathbf{F}$ and $\mathbf{G}$, we can estimate the predicted semantic vectors

$$\hat{\mathbf{S}} = \mathbf{CF}$$

and the predicted form vectors

$$\hat{\mathbf{C}} = \mathbf{SG},$$

with the hat on **S** and **C** indicating that the predicted matrices are estimates (in the statistical sense) that approximate the gold standard vectors but will usually not be identical to these vectors. It is often convenient to focus on individual words, in which case we have that

$$\begin{aligned} \hat{\mathbf{s}} &= \mathbf{cF} \\ \hat{\mathbf{c}} &= \mathbf{sG}. \end{aligned}$$

To evaluate the accuracy of a comprehension mapping, the predicted semantic row vectors $\hat{\mathbf{s}}$ of $\hat{\mathbf{S}}$ are correlated with all the corresponding semantic row vectors in the gold standard semantic matrix **S**. If the predicted semantic vector of a word form is closest to its target vector, it is counted as correct. This accuracy measure is referred to as correlation accuracy in the following, and we will sometimes denote it as `accuracy@1`[1]. More lenient accuracy measures `accuracy@k` accept model performance as satisfying when $\hat{\mathbf{s}}$ is among the top k nearest semantic vectors of the gold standard semantic vector **s**. For detailed introductions to the DLM, see Baayen et al. (2018a, 2019); Heitmeier et al. (2021, 2023).

3 Data

For the present study, we used a subset of 2646 singular and plural Dutch nouns and verbs (for which frequency was at least 1) taken from a dataset originally extracted from the Dutch CELEX database (Baayen et al., 1995) by Ernestus and Baayen (2003)[2]. These words have monomorphemic stems ending in an obstruent that is realized as voiceless when word-final but that in syllable onset appears with voicing in some words and without voicing in others. We used 300-dimensional Dutch fasttext embeddings (Grave et al., 2018) to represent semantics[3]. Since we were unable to obtain word embeddings for all words in our dataset, this left us with 2638 word forms (i.e. excluding 8 word forms). Because this dataset is quite small, we also present results with a larger dataset extracted from the Dutch Lexicon Project (Keuleers et al., 2010) in later sections. The frequencies of use that we use in this study are taken from CELEX. In what follows, we first present results that make use of bi-grams for representing word forms (used in their orthographic representation), resulting in a dimensionality of 360. The use of bigrams is motivated by the choice to minimize the carbon footprint of our simulations. However, for the new, computationally highly efficient frequency-sensitive model that is at the heart of this study, we made use of trigrams.

In what follows, for expositional simplicity, we focus mainly on comprehension. We note, however, that frequency-informed mappings are equally important for modeling production.

4 Endstate Learning

The first implementation of the DLM in the R (R Core Team, 2020) package **WpmWithLdl** (Baayen et al., 2018b) estimates the 'endstate' of learning. This implementation constructs a mapping between the cue matrix **C** and the semantic matrix **S** using the pseudo-inverse. Computing the pseudo-inverse as implemented in **WpmWithLdl** is expensive and prohibitively so for larger datasets. Fortunately, there now exists a very efficient method implemented in the Julia (Bezanson et al., 2017) package **JudiLing**[4] that makes use of the Cholesky decomposition (Luo, 2021; Heitmeier et al., 2023). This, together with additional speed-ups due to Julia being in general faster than R and the use of sparse matrices means that **JudiLing** can handle much larger datasets compared to **WpmWithLdl** (Luo, 2021).

[1] As only 2.7% of the words in our Dutch dataset are homographs, for simplicity we use 'strict evaluation' throughout this study, i.e. we do not take homographs into account while evaluating accuracy (further information on the various methods to evaluate accuracy in Heitmeier et al., 2021).

[2] Data, code and statistical models presented in this study are available at `https://osf.io/h2szj/`.

[3] `https://fasttext.cc/docs/en/crawl-vectors.html#models`

[4] `https://github.com/MegamindHenry/JudiLing.jl`

The endstate learning (EL) results in optimal mapping matrices that reduce the error between the predicted and the target vectors as much as possible, for all word forms. It is optimal in the least-squares sense, and the underlying mathematics are identical to that of multivariate multiple regression. This method is characterized as estimating the 'endstate' of learning, because the mappings it estimates can also be approximated by using incremental learning (WHL) applied to an infinite number of passes through one's dataset (assuming that each word occurs a single time in the dataset). With infinite experience, every word has been experienced an equal and infinite amount of times. Any effects of frequency of occurrence in this model are an epiphenomenon of lexical properties such as word length and neighborhood density (Nusbaum, 1985; Baayen, 2001).

Figure 1 illustrates the frequency-free property of EL. It presents the results obtained with an EL model for the dataset introduced in the preceding section, modeling comprehension with the matrix **F**. The average correlation accuracy for this model is low, at 40.8%; below, we will show how this accuracy can be improved to 83% by increasing the dimensionality of the form vectors.

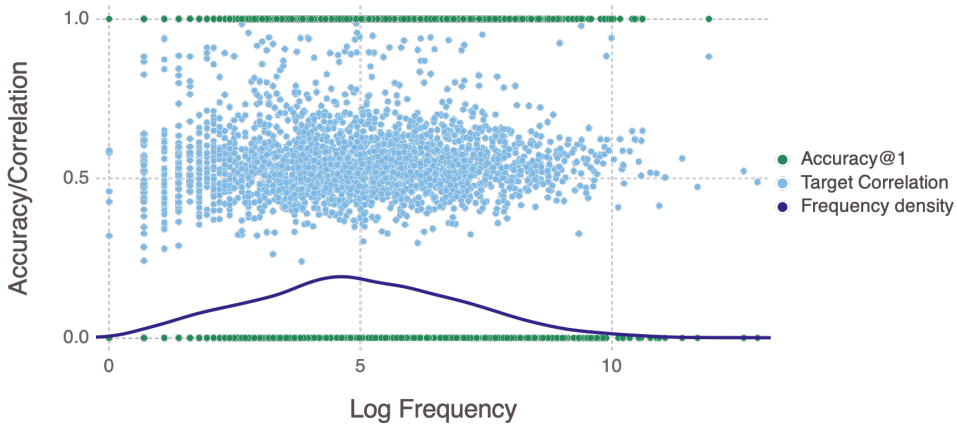


Figure 1: Endstate learning. The green dots on the horizontal lines at 0 and 1 represent the correlation accuracies@1 for the individual words (counted as correct if the semantic vector most correlated with the predicted semantic vector is the target), and the light blue dots represent the correlation values of words' predicted semantic vectors with their target vectors. The blue line presents the estimated kernel density for log frequency. Target correlation and accuracy are independent of (log) frequency.

To illustrate the absence of a relationship between word frequency and correlation accuracy, we used a binomial generalised linear model[5] with a logit link function, modeling the probability of correct recognition as a function of log frequency. Log frequency was not a good predictor of accuracy ($p > 0.05$).

In summary, one important advantage of endstate learning is that it is extremely efficient. A second important advantage of this method is that it is frequency-free: it shows what can be learned when frequency of use is not allowed to play a role. In other words, the EL probes the quantitative structure of a dataset, under the assumption that usage can be ignored. Thus, the EL method dovetails well with generative approaches to morphology that deny any role to usage in grammar and work with types only.

However, this is also the achilles heel of the EL method: EL is purely type-based and is blind to the consequences of token frequency for learning. Well-established effects of frequency of use (see, e.g., Baayen et al., 1997; Bybee and Hopper, 2001) are not captured. This is why an alternative way of estimating mappings between form and meaning is implemented in the **JudiLing** package: incremental learning.

[5] `https://github.com/JuliaStats/GLM.jl`

5 Incremental learning

Instead of making use of the efficient method estimating the endstate of learning, we can also learn the mappings incrementally using the Widrow-Hoff learning rule (Widrow and Hoff, 1960), a form of error-driven learning closely related to the Rescorla-Wagner learning rule (Rescorla and Wagner, 1972; for a discussion of the Widrow-Hoff learning rule in language research see also Milin et al., 2020). Here, the idea is that each time a word is encountered, a learning event occurs, and the mappings between form and meaning are updated in such a way that next time the same word is encountered (if no unlearning occurs intermittently) the mappings will be more accurate. How fast learning takes place is controlled via the learning rate parameter η. High learning rates lead to fast learning and unlearning, whereas low learning rates result in slower learning.

Incremental learning has the advantage that we can model learning in a frequency-informed way by translating frequencies into learning events. For example, if a word has a frequency of 100, it is presented for learning 100 times. We used the Widrow-Hoff learning rule as implemented in **JudiLing** to incrementally learn the mapping matrix $\mathbf{F}$. This mapping is now frequency-informed. We experimented with three different learning rates: 0.01, 0.001, and 0.0001. The average correlation accuracy for the three simulations was 15.5%, 14.0% and 10.1% respectively. As can be seen in the left panel of Figure 2, when learning the mappings incrementally according to their frequency, a clear relationship between frequency and accuracy is obtained for all learning rates: the more frequent a word is, the more accurately it is learned. A binomial GLM indicated a highly significant relationship between frequency and accuracy ($p < 0.001$ for all learning rates). The lower the learning rate, the steeper the increase in accuracy: accuracy is lower for low-frequency words and higher for high-frequency ones.

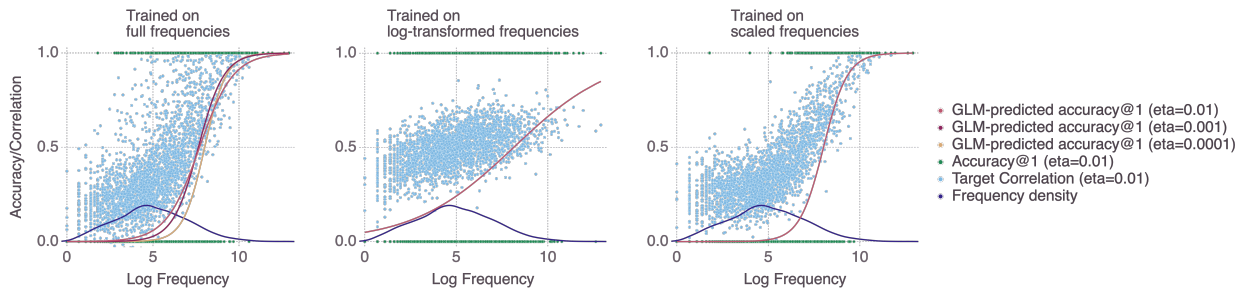


Figure 2: Relationship between accuracy and frequency for incremental learning. **Left panel:** Mapping trained using full frequencies. Predicted accuracy is depicted for three different learning rates ($\eta \in (0.01, 0.001, 0.0001)$), and the light blue dots present target correlations for $\eta = 0.01$. **Center panel:** Mapping trained using log-transformed frequencies. **Right panel:** Mapping trained using frequencies scaled by a factor of 100.

There are two disadvantages to this approach, one practical, and the other theoretical. The theoretical problem is that for many datasets, there is no intrinsic order in which words are learned. For the present dataset, which is basically a word frequency list, we do not have available any information about the order in which Dutch speakers encounter these words over their lifetime. It is only for intrinsically ordered data, such as child directed speech in corpora such as CHILDES (MacWhinney, 2000), that incremental learning comes into its own (see below). The practical problem is that updates with the Widrow-Hoff learning rule are computationally expensive.

For the present dataset, estimating the mapping matrix $\mathbf{F}$ took ≈25 minutes on a MacBook Pro (2017) with a 3.1 GHz Quad-Core Intel Core i7 processor, even though the use of bi-grams resulted in a form dimensionality that was already too small to obtain good accuracy. If we would use the better-suited tri-grams (see below for further details), the estimated time for computing the mapping matrix $\mathbf{F}$ increases strongly even in an optimized language such as Julia.

There are two obvious ways in which one might want to avoid the costly computations of WHL. One option is to transform frequencies by taking logarithms (see, e.g., Seidenberg and McClelland, 1989). The resulting average correlation accuracy is 26.1%. The relationship between frequency

and accuracy can be seen in the center panel of Figure 2. While the log transformation does indeed reduce computational costs, it is of course questionable whether such a transformation is justified and realistic. Low frequency words become proportionally more accurate, while high frequency ones become less so. If WHL with empirical frequencies for time-ordered learning events is taken as a gold standard, then a log-transformation distorts our estimates considerably.

A second option is to simply scale down frequencies by dividing them by a fixed number. By applying a ceiling function to the result, we avoid introducing zero frequencies. Training the model using frequencies divided by 100 speeds up the learning to ≈12 seconds and does not distort the learning curve (see right panel of Figure 2). The disadvantage of this method is that there are far fewer learning events. As a consequence, words are learned less well. Accordingly, the average correlation accuracy drops to 10.3%.

In summary, it is in principle possible to estimate mapping matrices with incremental learning. This is theoretically highly attractive for data that are intrinsically ordered in learning time (see, e.g., Heitmeier et al., 2022, for the modeling of within-experiment learning). For unordered data, some random order can be chosen, but for larger datasets, it would be preferable to have a method that is agnostic about order but nevertheless accounts in a principled way for the consequences of experience for discriminative learning.

6 Non-incremental, yet frequency-informed mappings

A solution to this conundrum is to construct frequency-informed mappings between form and meaning. Thinking back to incremental learning, learning a word w_i with frequency count f_i involved learning the mapping from a cue vector $\mathbf{c}_i$ to the word's meaning $\mathbf{s}_i$ f_i times. We could thus construct matrices $\mathbf{C}_f$ and $\mathbf{S}_f$ reflecting the entire learning history: $\mathbf{C}_f$ and $\mathbf{S}_f$ are $\mathbf{C}$ and $\mathbf{S}$ with word forms w_i and semantic vectors $\mathbf{s}_i$ repeated according to their frequency count f_i. We are looking for the mapping $\mathbf{F}_f$ and $\mathbf{G}_f$ such that

$$\mathbf{S}_f = \mathbf{C}_f \mathbf{F}_f$$

$$\mathbf{C}_f = \mathbf{S}_f \mathbf{G}_f$$

Formally, let $\mathbf{C} = \begin{pmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \dots & \mathbf{c}_m \end{pmatrix}^T \in \mathbb{R}^{m \times r}$ and $\mathbf{S} = \begin{pmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \dots & \mathbf{s}_m \end{pmatrix}^T \in \mathbb{R}^{m \times q}$ where each word w_i of the m wordforms corresponds to a row in the two matrices with cue vector $\mathbf{c}_i$ and semantic vector $\mathbf{s}_i$. Each word form w_i has a frequency count f_i.

We can create two new matrices $\mathbf{C}_f$ and $\mathbf{S}_f$ where the cue and semantic vectors of the wordforms are repeated according to their frequency count f_i. We want to find the mapping matrix $\mathbf{F}_f$ mapping from $\mathbf{C}_f$ to $\mathbf{S}_f$. We use a closed-form solution for computing the mapping matrix (see also Baayen et al., 2018a):

$$\mathbf{F}_f = (\mathbf{C}_f^T \mathbf{C}_f)^{-1} \mathbf{C}_f^T \mathbf{S}_f \tag{1}$$

$$\overset{\text{see Supplementary}}{=} \left(\sum_{i=1}^{m} f_i \mathbf{c}_i \mathbf{c}_i^T \right)^{-1} \left(\sum_{i=1}^{m} f_i \mathbf{c}_i \mathbf{s}_i^T \right) \tag{2}$$

$$= \left(\sum_{i=1}^{m} \sqrt{f_i} \mathbf{c}_i \left(\sqrt{f_i} \mathbf{c}_i \right)^T \right)^{-1} \left(\sum_{i=0}^{m} \sqrt{f_i} \mathbf{c}_i \left(\sqrt{f_i} \mathbf{s}_i \right)^T \right) \tag{3}$$

$$\overset{\text{see Supplementary}}{=} \left(\sum_{i=1}^{m} \sqrt{\frac{f_i}{k}} \mathbf{c}_i \left(\sqrt{\frac{f_i}{k}} \mathbf{c}_i \right)^T \right)^{-1} \left(\sum_{i=1}^{m} \sqrt{\frac{f_i}{k}} \mathbf{c}_i \left(\sqrt{\frac{f_i}{k}} \mathbf{s}_i \right)^T \right) \tag{4}$$

with a constant $k \in \mathbb{R}_{>0}$. Since k does not change the solution, we can set it such that the algorithm is numerically more stable, for example to $k = \max_{j \in 1:m} f_j$ so that we have $p_i = \frac{f_i}{\max_{j \in 1:m} f_j}$ and therefore

$$\mathbf{F}_f = \left(\sum_{i=1}^{m} \sqrt{p_i} \mathbf{c}_i \left(\sqrt{p_i} \mathbf{c}_i \right)^T \right)^{-1} \left(\sum_{i=1}^{m} \sqrt{p_i} \mathbf{c}_i \left(\sqrt{p_i} \mathbf{s}_i \right)^T \right) \tag{5}$$

Now let $\mathbf{P} \in \mathbb{R}^{m \times m}$ be a diagonal matrix with $p_{ii} = p_i$ for $i \in 1, ..., m$. Then we can define $\tilde{\mathbf{C}} = \sqrt{\mathbf{P}}\mathbf{C}$ and $\tilde{\mathbf{S}} = \sqrt{\mathbf{P}}\mathbf{S}$ so that $\tilde{\mathbf{c}}_i = \sqrt{p_i}\mathbf{c}_i$ and $\tilde{s}_i = \sqrt{p_i}\mathbf{s}_i$. Then we have

$$\begin{aligned}\mathbf{F}_f =& \left(\sum_{i=1}^{m} \tilde{\mathbf{c}}_i \tilde{\mathbf{c}}_i^T\right)^{-1} \left(\sum_{i=1}^{m} \tilde{\mathbf{c}}_i \tilde{\mathbf{s}}_i^T\right) \\ =& (\tilde{\mathbf{C}}^T \tilde{\mathbf{C}})^{-1} \tilde{\mathbf{C}}^T \tilde{\mathbf{S}}\end{aligned}$$

Therefore, the pair $(\tilde{\mathbf{C}}, \tilde{\mathbf{S}})$ has the same mapping matrices as $(\mathbf{C}_f, \mathbf{S}_f)$.

Practically, this means that we can first weigh $\mathbf{C}$ and $\mathbf{S}$ with the pertinent frequencies to obtain $\tilde{\mathbf{C}}$ and $\tilde{\mathbf{S}}$. We can then use the closed-form solution (making use of algorithms such as Cholesky decomposition) to obtain frequency-informed mappings between these two matrices.[6]

In what follows, we sketch the new possibilities enabled by this method, to which we will refer as frequency-informed learning (FIL). A first, practical, advantage of FIL is that it is efficient and fast. A second, theoretical, advantage is that predictions are available for datasets for which no information about the order of learning is available.

6.1 Low-dimensional modeling

We first consider modeling studies using the low-dimensional vectors that we used in the preceding sections. The low dimensionality of these vectors is motivated by the wish to avoid long computation times for WHL for these exploratory studies.

Figure 3 shows the relationship between log-transformed frequency and accuracy predicted by a logistic GLM regressing the correctness of FIL responses on log frequency. `Accuracy@k` is set to 1 if a word's predicted semantic vector is among the k predicted vectors that are most correlated with the targeted semantic vector, and to 0 otherwise.

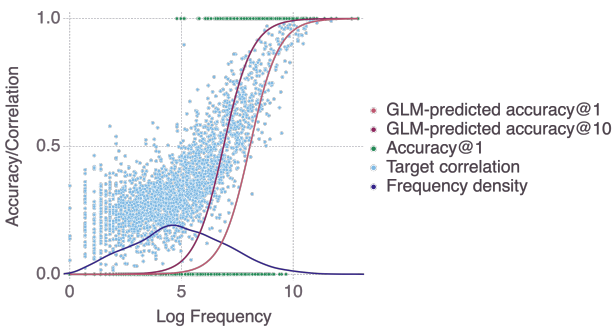


Figure 3: Frequency-informed learning. The red line presents the model's predictions when a success is defined as the predicted vector being the closest to its gold standard target vector in terms of correlation (`accuracy@1`). The purple line represents model predictions when a success is defined as the correlation being among the top 10 (`accuracy@10`). The blue line visualizes the estimated density of the log-transformed frequencies. The green dots represent the successes and failures for `accuracy@1`. The light blue dots represent for each word the correlation of the predicted and gold-standard semantic vectors.

Correlation and `accuracy@1` increase for higher frequency, as required. When comparing accuracies with the frequency distribution depicted in the same plot, we can also see that there is a large number of low frequency words with very low (predicted) accuracy. Although for most words, the correlations are relatively high, the overall `accuracy@1` is low, at 9.9%. When we relax our criterion for accuracy, use `accuracy@10`, counting a predicted vector as correct if the target vector

[6]Note that while we here work with a learning rule for continuous vectors, FIL is also applicable to mappings between discrete vectors, as in Rescorla-Wagner learning (Rescorla, 1967) which is used in Naive Discriminative Learning (NDL, Baayen et al., 2011): instead of having to impose some random order on the learning events, the expected value irrespective of order can be estimated using FIL.

is among the 10 closest semantic vectors, we see that the accuracy starts to rise earlier, but there is still a significant portion of words for which even `accuracy@10` is zero.

The relatively large number of words with very low accuracy raises the question of whether accuracy can be improved by using log-transformed frequencies for FIL. Figure 4 clarifies that accuracies increase for lower-frequency words, but decrease somewhat for higher-frequency words. The average `accuracy@1` is accordingly higher at 34.3%.

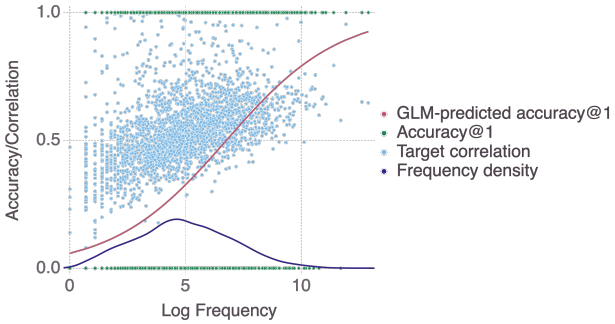


Figure 4: `Accuracy@1` as a function of log frequency, using frequency-informed learning with log-transformed frequencies.

The upper half of Table 1 provides an overview of the `accuracies@1` for different combinations of learning (incremental/frequency-informed) and kind of frequency used (untransformed, scaled, or log-transformed). Here, we observe first of all that endstate learning offers the highest accuracy (40.8%), followed by log-frequency informed learning (34.3%).

low-dimension simulations		
Model	Average `accuracy@1`	Frequency-weighted `accuracy@1`
Endstate learning (EL)	40.8%	32.0%
Incremental learning ($\eta = 0.01$)	15.5%	78.1%
Incremental learning ($\eta = 0.001$)	14.0%	79.1%
Incremental learning ($\eta = 0.0001$)	10.1%	75.6%
Incremental learning log-frequencies	26.1%	57.6%
Incremental learning scaled frequencies	10.3%	75.3%
Frequency-informed learning (FIL)	9.9%	74.9%
Log-frequency-informed learning	34.3%	71.2%
high-dimension simulations		
Model/Dataset size	Average `accuracy@1`	Frequency-weighted `accuracy@1`
Endstate learning (EL)/Small	83.0%	79.8%
Frequency-informed learning (FIL)/Small	22.3%	89.4%
Endstate learning (EL)/Large	67.8%	43.7%
Frequency-informed learning (FIL)/Large	5.1%	79.8%

Table 1: Comparison of average and frequency-weighted `accuracy@1` (the term "frequency-weighted accuracy" is introduced in Section 6.4) across simulation studies. The small dataset contained 2,638 word forms, the large dataset 13,669.

Figure 5 highlights the differences between the model set-ups, comparing with FIL the effect of a log-transformation (left panel), of scaling (center panel), and of the learning rate (right panel). It is noteworthy that frequency-informed learning with log-transformed frequencies departs the most from both FIL and incremental learning, which suggests that training on log frequency may artefactually increase learning performance.

Secondly, it can be observed that the incremental learning based on scaled frequencies is closest to frequency-informed learning in terms of average `accuracy@1`, as well as to incremental learning with the lowest learning rate. This suggests a) that scaling frequencies has a similar effect as lowering

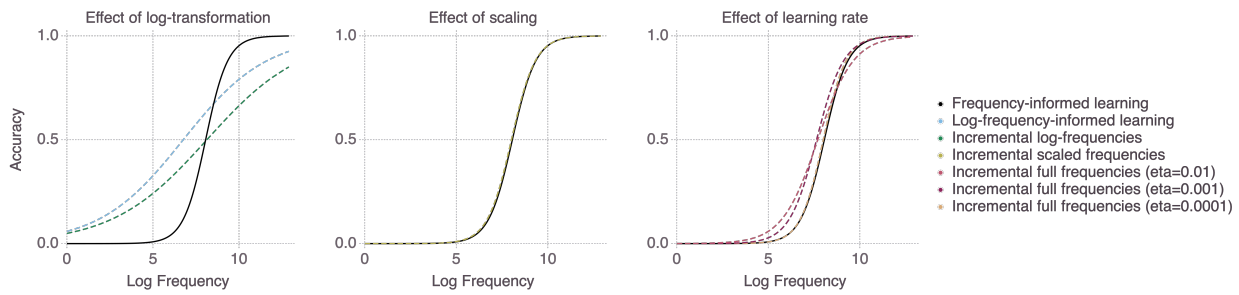


Figure 5: Comparison of methods. GLM-predicted `Accuracy@1` with frequency-informed learning is plotted as a black solid line: The left panel compares methods based on log-frequencies, the center panel compares methods based on scaled frequencies and the right panel compares incremental learning with different learning rates. Incremental learning with scaled frequencies or with a very low learning rate ($\eta = 0.0001$) is closest to frequency-informed learning.

the learning rate in incremental learning and b) that frequency-informed learning approximates incremental learning for very low learning rates.

6.2 High-dimensional modeling

Importantly, the mappings that we used thus far are suboptimal: the dimensionality of the semantic vectors was small and the use of bi-grams for the form vectors often shows underwhelming performance (see Heitmeier et al., 2021, for further discussion of the underlying reasons). While opting for low dimensionality decreased the computational costs for incremental learning immensely and was therefore necessary for comparing methods, we now proceed to investigate the accuracy of frequency-informed learning for larger, more discriminative cue matrices. To this end, we employed letter trigrams resulting in a form vector dimensionality of 1776 (compared to 360 previously).

The model for the endstate of learning now performs much better, at an `accuracy@1` of 83% instead of 40.8%. In other words, with infinite experience of just the words in this dataset, and with all token frequencies going to infinity, this is the best our simple multivariate multiple regression approach can achieve (conditional on the way in which we encoded form and meaning).

A model using FIL obtained an average `accuracy@1` of 22.3%, which is clearly superior to the 9.9% obtained for the lower-dimensional model. The upper panel of Figure 6 presents the predicted accuracy curves for the high-dimensional FIL model in red, and the low-dimensional FIL model in black. We see a rise in accuracy for lower frequencies.

6.3 Increasing the dataset size

Having established the relationship between EL, WHL and FIL, we can also investigate how EL and FIL fare when the modelled dataset is significantly larger. To this end we used all 13,669 words from the Dutch Lexicon Project (DLP; Keuleers et al., 2010) for which we were able to obtain fasttext embeddings, and modelled the dataset using trigram representations (resulting in a form vector dimensionality of 4,678) as well as frequencies extracted from CELEX (as included in the DLP). We found that accuracies in general were clearly lower: For EL, the `accuracy@1` was 67.8% and for FIL it was 5.1%. Qualitatively, the differences between EL and FIL are therefore similar (see also Figure 6), but our simple linear mappings clearly perform less well with very large datasets, especially when taking into account frequency.

6.4 Frequency-weighted accuracy

A final question is whether the way we have been evaluating accuracy so far is reasonable. In the average `accuracy@1` each word's accuracy contributes the same, that is, we have effectively calculated accuracy across all word types in our corpus. However, it can be argued that comprehending

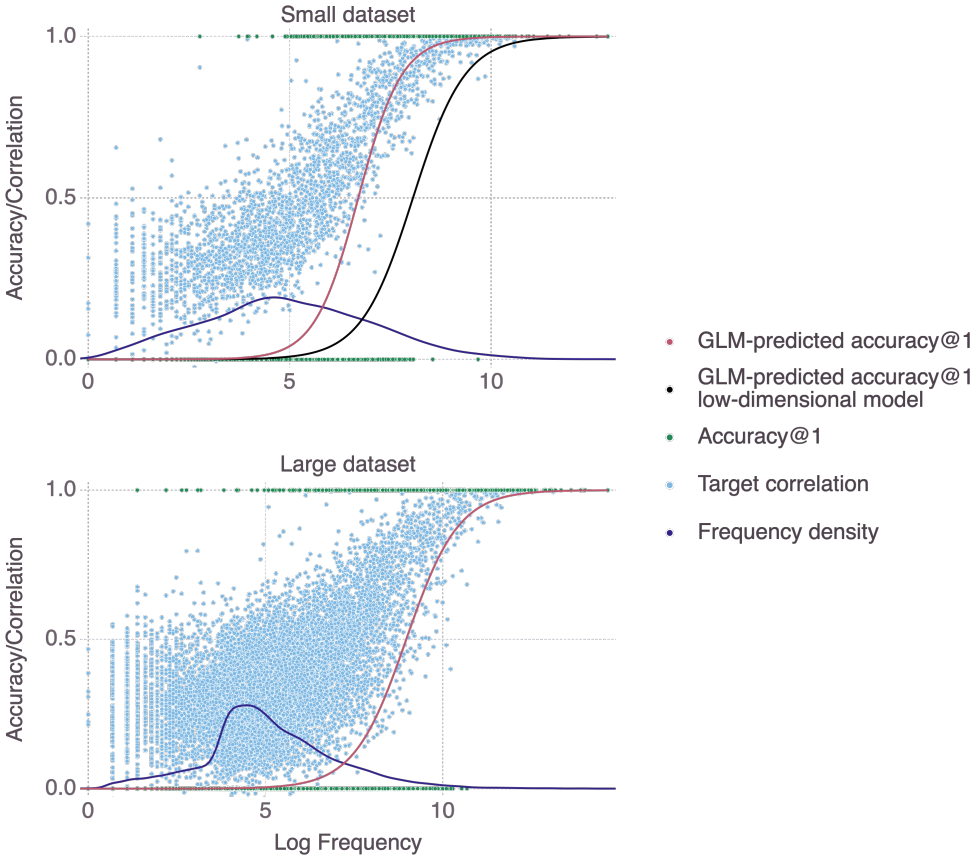


Figure 6: Predicted `accuracy@1` as a function of log frequency for a high-dimensional representations of form and meaning (red line). The black line shows the predicted accuracy based on the low-dimensional model, for comparison. The light blue dots represent the target correlations in the high-dimensional model. The small dataset refers to the dataset with 2,638 word forms based on Ernestus and Baayen (2003), the large dataset to the dataset created from the DLP (Brysbaert and New, 2009) including 13,669 word forms.

high frequency words is much more important than low frequency words, and that we should be calculating accuracy across word tokens instead. Practically this would mean that if we would go through a corpus of written English, instead of counting how many unique words our model is able to comprehend, we would count how many of all of the encountered word tokens are understood correctly.

Following this line of argumentation we also provide frequency-weighted accuracies in Table 1. For instance, the word with the highest frequency in the Dutch Lexicon Project (large dataset) is *de* (eng. *the*). Since this word accounts for 7% of all word tokens in the Dutch Lexicon Project (calculated by summing up the frequencies of all word types), it also contributes 7% to our frequency-weighted accuracy measure. There are also 349 words with a frequency of 0 in CELEX (2.6% of all word types) which accordingly do not contribute to the frequency-weighted accuracy at all. We find that with this method, the results flip in comparison with average `accuracy@1`: Generally, FIL and WHL based on untransformed frequencies perform the best, followed by methods based on log-transformed frequencies, while EL clearly performs the worst. For example, for the high-dimensional simulations on the large dataset, EL has a frequency-weighted `accuracy@1` of 43.7%, while FIL achieves 79.8%.

To summarise, FIL provides an efficient way of estimating mappings between frequency-informed representations of form and meaning. FIL does not reach the average accuracy across types of EL, but, importantly, unlike for EL, accuracy varies systematically with frequency in a way similar to how human accuracy is expected to vary with frequency. Moreover, FIL clearly outperforms EL

when accuracy is calculated across tokens rather than types. The question addressed in the next section is whether FIL indeed provides predictions that match well with a particular behavioral measure: reaction times in a visual and an auditory lexical decision task.

7 FIL-based modeling of reaction times

7.1 Visual word recognition in Dutch

In order to assess the possibilities offered by FIL-based modeling (using untransformed frequencies and the high-dimensional form and meaning vectors) for predicting behavioral measures of lexical processing, we return to the large dataset of reaction times to Dutch words described in Section 6.3 (13,669 words represented as trigrams with a form dimensionality of 4,678, and semantics represented using 300-dimensional fasttext vectors).

Figure 7 presents the partial effects according to three Gaussian Location-Scale Generalised Additive Models (Wood, 2011, GAMs,) fitted to the response latencies in the Dutch Lexicon Project. Response latencies were inverse transformed (-1000/RT) in order to avoid marked non-normality in the residuals. (Effects for the untransformed RTs are very similar in shape, but confidence intervals are not reliable due to the marked departure from normality of the residuals). The left-hand panels present the partial effects for the mean (in blue), the right panels the partial effects for the variance (in green, on the $\log(\sigma - 0.01)$ scale). The upper panels pertain to a GAM predicting RT from log frequency (AIC -16384.8). Mean and variance decrease non-linearly with increasing frequency. The smooth for the mean shows the kind of non-linearity that is typically observed for reaction time data (see, e.g., Baayen, 2005; Miwa and Baayen, 2021): the effect of log frequency levels off strongly for high-frequency words and to a lesser extent also for low-frequency words. The partial effect of the variance is less wiggly and decreases as frequency increases. This decrease in variability for increasing frequency has at least two possible sources. First, high-frequency words are known to all speakers, whereas low-frequency words tend to be specialized and known to smaller subsets of speakers. Second, more practice, as in the case of high-frequency words, typically affords reduced variability (see, e.g., Tomaschek et al., 2018).

A model-based measure that we expected to correlate with reaction time is the proximity of a word's predicted semantic vector to its corresponding gold standard vector. The more similar the two vectors are, the better a word's meaning is reconstructed from its form. In other words, the more effective a word form is in approximating its meaning, the more word-like it is and the faster a lexicality decision can be executed. We used the correlation r of $\hat{\mathbf{s}}$ and $\mathbf{s}$ as a measure of semantic proximity. Since for large r, lexical decision times are expected to be short, whereas for small r long decision times are more likely, we took $1 - r$ as a measure that we expect to enter into a positive correlation with RT. This measure, when based on FIL, has a density that is roughly symmetrical, and that does not require further transformations to avoid adverse affects of outliers.

The panels in the middle row present the partial effect of $1 - r$ as predictor of RT using FIL (AIC -12900.75), and the bottom panels present the corresponding partial effects using EL (AIC -10896.19). The GAM with the FIL-based predictor clearly provides the superior fit to the observed response latencies. The effect of $1 - r$ on the mean is fairly linear for FIL but U-shaped for EL. A strong effect on the variance is present for FIL, but absent for EL. Effects are opposite to those of frequency, as expected, as $1 - r$ is constructed to be positively correlated with RT. The absence of a highly significant effect on the variance in RT for EL ($p = 0.0356$) is perhaps unsurprising given that EL learns the mapping from form to meaning to perfection (within the constraints of a linear mapping and the type distributions of forms and meanings), and hence little predictivity for human processing variance is to be expected. The question of why $1 - r$ has a U-shaped effect on RT for EL will be addressed below.

In summary, FIL generates a predictor $(1-r)$ that is better aligned with observed RTs. However, log frequency provides a better fit (AIC -16384.8). Here, it should be kept in mind that word frequency is correlated with many other lexical properties (Baayen, 2010), including word length and number of senses. Longer, typically less frequent, words require more fixations, and hence are expected to have longer reaction times. The greater number of senses for higher-frequency

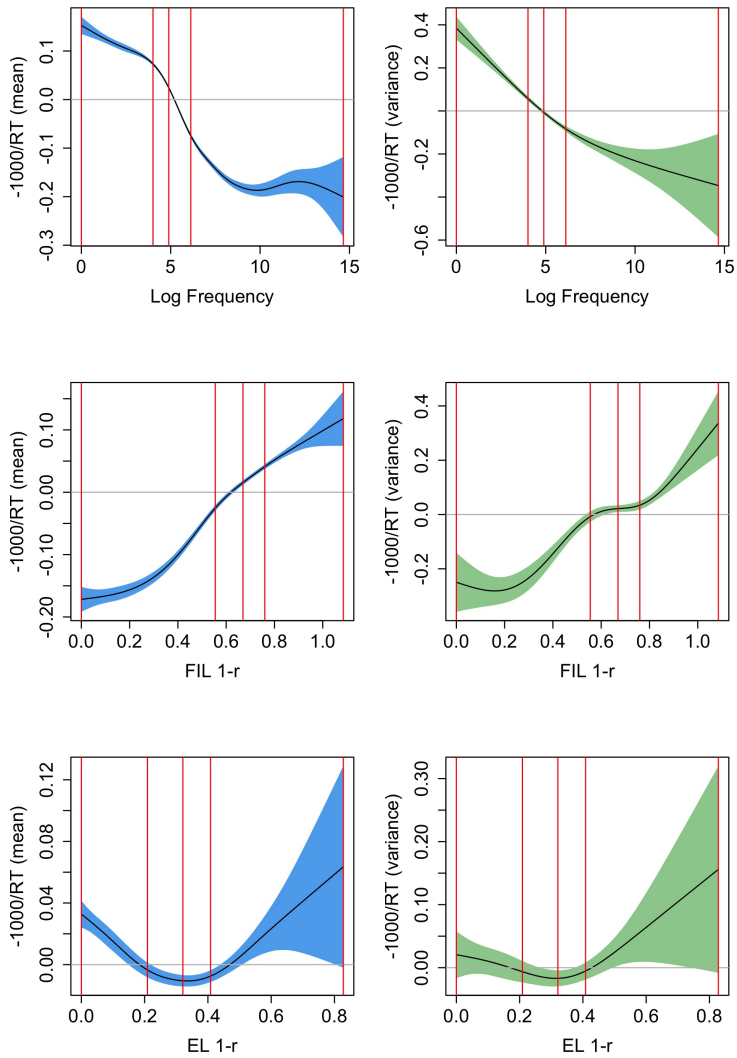


Figure 7: Partial effects for mean (left, confidence intervals in blue) and variance (right, confidence intervals in green, the y-axis is on the $\log(\sigma - 0.01$ scale), for Gaussian Location-Scale GAMs predicting reaction times from log frequency (upper panels), from $1 - r$ based on FIL (center panels), and from $1 - r$ based on EL (bottom panels). The vertical red lines represent the 0%, 25%, 50%, 75%, and 100% percentiles.

words are not directly reflected in the embeddings, which typically consist of a single embedding per unique word form, rendering the mapping less precise. As a consequence, there is necessarily imprecision in measures derived from our learning models. Furthermore, $1 - r$ is only one of the many learning-based measures that predict lexical decision times, see Heitmeier et al. (2022) for detailed discussion.

What we have not considered thus far is how $1-r$ is affected by frequency when learning is based on FIL and on EL, and how the effect of frequency on these measures compares to the effect of frequency on reaction times. Figure 8 presents the partial effects of Gaussian Location-Scale GAMs for FIL (upper panels) and EL (lower panels). The leftmost panels present effects for the mean, and the center panels effects for the variance. For FIL, the highest-frequency words are learned to perfection, and hence the variance in $1 - r$ is extremely small. To highlight the variance function for most of the datapoints, the upper right panel restricts the y-axis to a smaller range. For all points in the interquartile range of frequency, the variance increases with frequency.

Comparing the partial effects for the means, FIL presents a curve that is similar to the partial effect of log frequency on reaction time (see Figure 7, upper left panel), whereas the curve for

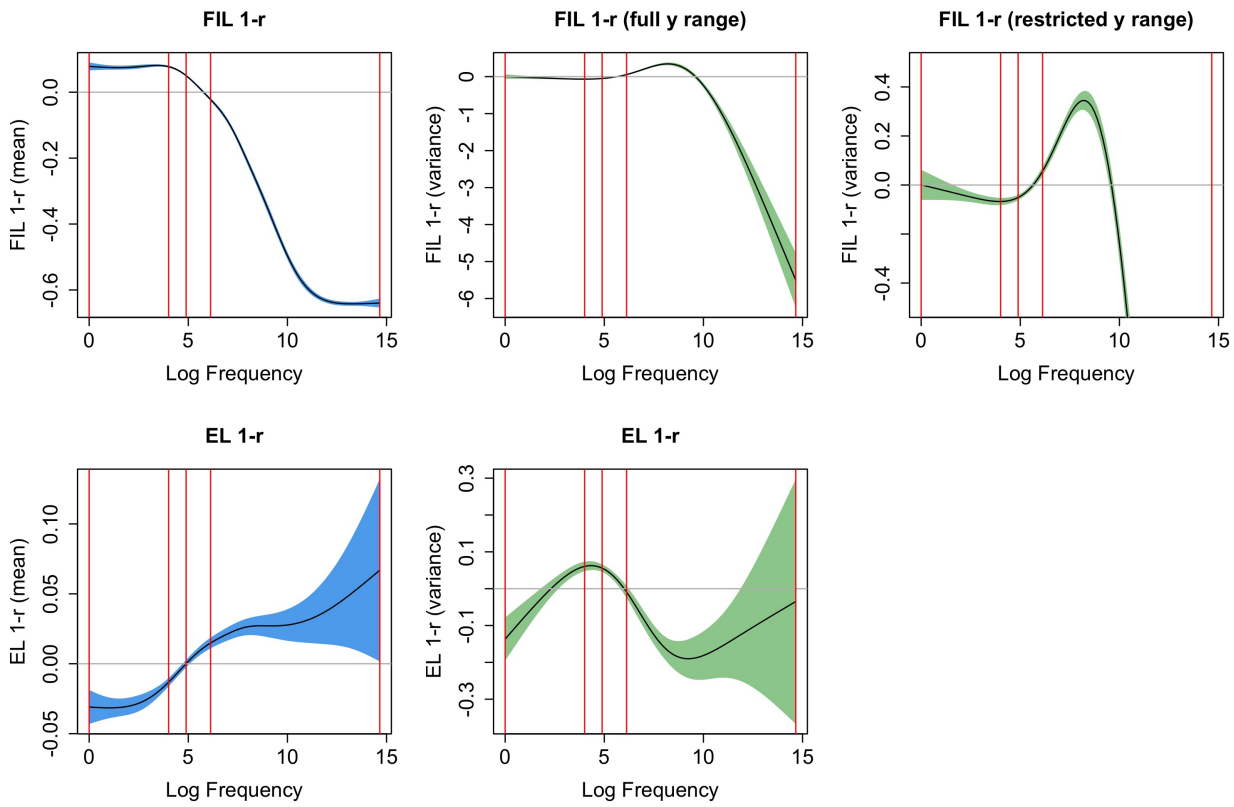


Figure 8: Partial effects for mean (left, confidence intervals in blue) and variance (right, confidence intervals in green, the y-axis is on the $\log(\sigma - 0.01$ scale), for Gaussian Location-Scale GAMs predicting $1 - r$ from log frequency for FIL (upper panels), and for EL (bottom panels). The panel in the upper right zooms in on the partial effect of variance shown to its left. The vertical red lines represent the quartiles of log frequency.

EL is an increasing function of frequency, rather than a decreasing function. The reason for this is straightforward: higher-frequency words share more segments and substrings with other words than lower frequency words, they are shorter, and tend to have more form neighbors (Nusbaum, 1985; Baayen, 2001). As a consequence, they provide less information for the mapping from form to meaning, resulting in less accurate predicted semantic vectors, and hence higher values of $1 - r$. This disadvantage of being shorter and easier to pronounce is overcome in FIL. FIL, and also incremental learning, provide higher frequency words with more learning opportunities compared to lower-frequency words.

The U-shaped curve of $1 - r$ using EL as predictor of reaction time (see the lower left panel of Figure 7) can now be understood. For EL, median $1-r$ is 0.32, which is where the curve reaches its minimum. As we move to the left of the median, word frequency goes down, and as a consequence, salient segments and segment combinations (cf. English *qaid*, 'tribal chieftain', which has the highly infrequent bigram *qa*) are more common. These salient segments and n-grams allow these words to map precisely onto their meanings, much more so than is warranted by their very low frequencies of use. Although EL provides estimates of $1 - r$ that are low, EL underestimates the difficulty of learning these words in actual usage. As a consequence, actual reaction times are higher than expected given the computed $1-r$. Conversely, when we move from the median to the right, we see the expected slowing due to being further away from the semantic target. Apparently, the greater form similarity and denser form neighborhoods that characterize higher-frequency words results in estimates of $1-r$ that are reasonably aligned with reaction times, albeit by far not as well as when FIL is used to estimate $1 - r$.

We have seen that the variance in RTs goes down as word frequency is increased, which we attributed to higher frequency words being known and used by more speakers than is the case for

Prime	Target	Condition
chun1	**chun1**	ST
chun3	**chun**1	S
tong**1**	chun**1**	T
liao2	chun1	UR

Table 2: The design of priming conditions for Experiments 1 and 2 of Lee (2007). For the target word *chun1* 'spring', the prime either shares both syllable and tone with it (ST), or only syllable (S) or only tone (T). In the UR condition, neither syllable nor tone is the same. The parts shared between primes and targets are marked in bold.

lower-frequency words, and the general reduction in variability that comes with increased practise. These kinds of factors are not taken into account in the current FIL mapping. It is therefore interesting to see that without these factors, FIL suggests that, at least for the interquartile range of frequency, the variance increases with frequency. But why this might be so is unclear to us.

Considered jointly, these results provide good evidence that FIL adequately integrates frequency into linear discriminative learning, outperforming endstate learning by a wide margin, both qualitatively and quantitatively.

7.2 Spoken word recognition in Mandarin

Mandarin, as a tone language, alters pitch patterns to distinguish word meanings. There are four lexical tones in Mandarin: high level, rising, dipping, and falling, which will, for convenience, henceforth be referred to as T1, T2, T3, and T4 respectively. Take the syllable *ma* for example. It could mean 'mother', 'hemp', 'horse', or 'scorn', depending on which lexical tone it is associated with.

The role that tone plays in Mandarin word recognition has been widely discussed. Specifically, researchers are interested in whether native listeners exploit tonal information similarly as they do for segmental information. In other words, will a mismatch in tone (e.g., *ma3* and *ma1*) reduce the activation of a given word, to the same extent as a mismatch in segments (e.g., *ba1* and *ma1*)? Lee (2007) addresses this issue with an auditory priming experiment. In his study, four priming conditions were designed, as shown in Table 2. Among them, ST is an identity priming condition, and UR is a control priming condition. The experimentally critical conditions are S and T, where only either syllable or tone is shared between primes and targets. If tonal information is as important as segmental information in Mandarin, then the degree of priming should be similar for both conditions. On the contrary, differences should be observed if the two sources of information are treated differently by native listeners.

It was found that reaction times to the target words are shortest in the ST condition, hence most priming, as expected. Interestingly, mere tone sharing or syllable sharing is not sufficient to induce a reliable priming effect: RTs of both the S and T conditions do not differ significantly from the UR condition. But importantly, there is still a significant difference between the S and T conditions, where syllable sharing induces faster responses than tone sharing. In other words, more priming is found for syllable sharing than tone sharing. It is noteworthy that this pattern of results holds regardless of whether a long (250 ms, experiment 1) or short (50 ms, experiment 2) inter-stimulus interval is adopted in the experiment.

To model this priming experiment, we made use of the Chinese Lexical Database (Sun et al., 2018). In total 48,274 one- to four-character Mandarin words were selected, which include all the stimuli of the experiment, and for which fasttext word embeddings (Grave et al., 2018) are available. For cue representations, following Chuang et al. (2021), we created segmental and suprasegmental (tonal) cues separately. Thus, for a bisyllabic word such as *wen4ti2* 'question', the segmental cues will be triphones of `#we`, `wen`, `ent`, `nti`, `ti#`, and there will also be tritone cues of `#42` and `42#`. The separation of segmental and suprasegmental information, however, does not do justice to the fact that segments do have influence on tonal realizations and vice versa (e.g., Howie, 1974; Ho, 1976; Xu and Wang, 2001; Fon and Hsu, 2007). We therefore also made cues out of tone-

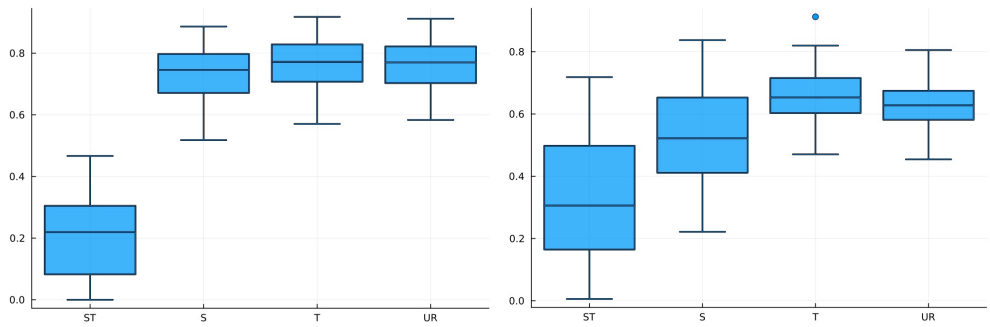


Figure 9: Boxplots of LDL simulated RTs for the four priming conditions in Lee (2007) with EL (left) and FIL (right).

segment combinations. To operationalize this, we marked tones on vowels, so that vowels with different tones are treated as separate phones. For the word *wen4ti2*, we then have additional tone-segment triphone cues of `#we4`, `we4n`, `e4nt`, `nti2`, `ti2#`. This resulted in an overall form vector dimensionality of 47,791.

We ran two LDL models, one with EL and the other with FIL. With EL, comprehension accuracy is at 83.71%. The accuracy is substantially worse with FIL; `accuracy@1` is 8.98%. As discussed previously, this is largely due to the low accuracy for especially low frequency words. After taking token frequency into account, the frequency-weighted accuracy is at 86.89%.

To model the RTs of the auditory priming experiment of Lee (2007), we calculated the correlation between the predicted semantic vector of the prime word ($\hat{s}_{prime}$) and the gold standard semantic vector of the target word (s_{target}), and again took $1 - r$ to predict RTs (see Baayen and Smolka, 2020, for the same implementation to simulate RTs for visual priming experiments). Results of simulated RTs with EL and FIL are presented in Figure 9. For EL (left panel), the simulated RTs of the ST condition is the shortest, unsurprisingly. For both the S and T conditions, the simulated RTs are similar to those of the UR condition, indicative of no priming. Tukey's HSD test reported no significant difference for any pairwise comparison among the S, T, and UR conditions. Although the general pattern of results is in line with the behavioral data, we however missed the crucial difference between the S and T conditions.

A different picture emerges with the FIL modeling. As shown in the right panel of Figure 9, not only does the ST condition induce faster responses than the other three conditions, but the simulated RTs of the S condition are also significantly shorter than those of the T condition ($p < .0001$, according to a Tukey's HSD test), as was found in the behavioral data. As this difference is absent in EL modeling, we conclude that FIL provides an estimate that better approximates the actual auditory comprehension of native Mandarin listeners.

8 But what about order?

Arguably, there is a piece of information missing when using FIL: order information. If a word is highly frequent early in a learning history and never occurs later, would it be forgotten by a WHL model but learned fairly well by a FIL model? To investigate how much of a problem this loss of order is in real-world longitudinal data, we used data of "Sarah" in the CHILDES Brown corpus (Brown, 1973) and of "Lily" in the PhonBank English Providence corpus (Demuth et al., 2006). To access all child-directed speech, we utilised the *childesr* package (Sanchez et al., 2019) to extract the data of all utterances not made by the target children themselves, resulting in 189,772 and 373,606 tokens respectively[7]. Of these, we kept all tokens for which pronunciations were available in CELEX (Baayen et al., 1995) and for which we could obtain word embeddings in 300-dimensional

[7]We ordered the data by age, followed by utterance id, followed by the token sequence within the utterance. Unfortunately, this part of *childesr* is not very well documented, so though sampling suggests that this corresponds to the original order of the data, we cannot be absolutely sure. However, the overall ordering by age should ensure that our results are valid even if some of the utterances might be in the wrong order.

Wikipedia2Vec word embeddings (Yamada et al., 2020). This resulted in 162,443 learning events (3,865 unique word tokens) for Sarah and 326,518 learning events (7,433 unique words tokens) for Lily. The cue matrix was created based on bigrams of CELEX DISC notation symbols, so for example, *thing* was represented as `#T`, `TI`, `IN`, `N#`, in order to model auditory comprehension (thus, the form vector dimensionality was 943 and 1,087 for Sarah and Lily respectively). We then trained the comprehension matrix $\mathbf{F}$ incrementally, evaluating after every 5,000 learning events the correlation of the predicted semantics of all word tokens with their target semantics, as well as keeping track of the frequency of each word token within the last 5,000 learning events. To gauge the effect of different learning rates we ran this simulation for $\eta \in \{0.1, 0.01, 0.001, 0.0001\}$.

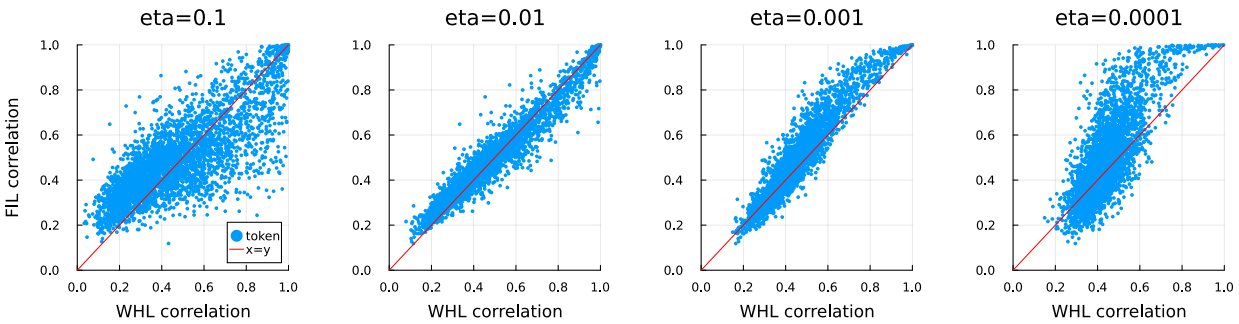
method	η	correlation accuracy@1		r(WHL, FIL)	
		Sarah	Lily	Sarah	Lily
FIL	—	12.9%	8%	—	—
WHL	0.1	**22.3%**	**13.8%**	.79	.73
	0.01	15.7%	10.9%	**.97**	.94
	0.001	6.9%	5.4%	.95	**.97**
	0.0001	2.1%	1.5%	.78	.80

Table 3: Correlation accuracies@1 for FIL and WHL with different learning rates (η). r(WHL, FIL) indicates the correlation between the target correlations obtained with WHL and the target correlations obtained with FIL (see Figure 10 for scatter plots). "Sarah" and "Lily" correspond to the child-directed speech in the Brown corpus (Brown, 1973) and Providence corpus (Demuth et al., 2006) respectively.

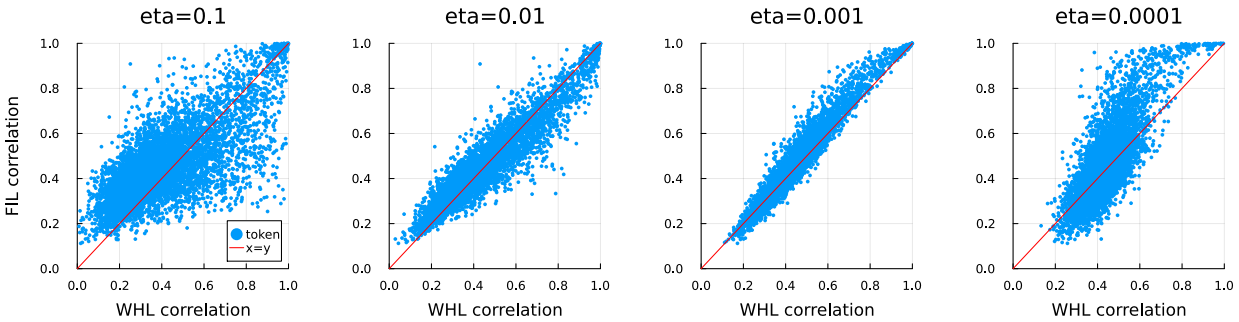
In order to investigate the effect of neglecting order during learning, we also trained a model with the same form and semantic matrix but using the FIL method. The FIL method results in a correlation accuracy of 12.9% for Sarah and 8.0% for Lily. For WHL the correlation accuracies vary across learning rates (Table 3), with accuracy decreasing for lower learning rates. FIL correlation accuracies are somewhat better than WHL accuracies for $\eta = 0.001$ and somewhat worse than for $\eta = 0.01$. Target correlations obtained with FIL and WHL are in general remarkably similar, correlated the highest for learning rates of 0.001 and 0.01 for "Lily" and "Sarah" respectively. This can also be observed visually in Figure 10: WHL and FIL target correlations are the least similar for $\eta = 0.1$ and $\eta = 0.0001$. Interestingly, for higher learning rates, low WHL correlations tend to be higher in FIL and vice versa, whereas for lower learning rates an advantage of FIL over WHL is more visible for higher accuracies, while low accuracies in WHL tend to be even lower in FIL.

Striking a balance between WHL accuracy and correlation between WHL and FIL accuracies, we now focus on $\eta = 0.01$. As can be seen in Figure 10 there are a few outliers where either FIL clearly outperforms WHL or vice versa. For the former case, the most apparent cases for Sarah are *no*, *mummy*, *cold*, *later*, *michael* and *told*. Except *no* which suffers due to its homophone *know*, all show a relatively higher frequency at the beginning of the learning trajectory compared to the end (see Figure 11a). Moreover, they tend to have overlapping di-phones: for example, *told* has overlap with *old* with which they are confused in the WHL model. Lily's data shows a similar pattern. WHL therefore unlearns when two conditions apply: a word occurs very infrequently in the later stages of the learning sequence and it has overlaps with cues in other words. WHL has an advantage over FIL when words have a higher frequency later in the sequence: for Sarah's data this is the case for *idea*, *old*, *mister*, *giraffe* and *alligator* (see Figure 11b).

To confirm this qualitative analysis quantitatively, we computed the mean, mode, skewness and kurtosis of the frequency distribution across time for each word. We then used these as predictors in individual (due to collinearity) Generalised Additive Models (GAMs; Wood, 2011) to predict the difference in target correlation in WHL and FIL for $\eta = 0.01$, i.e. predicting whether WHL or FIL perform better depending on the frequency distribution of a word across time (details on method in Supplementary Materials). We found that the higher the mean and mode, and the more negative the skew of the frequency distribution (all indicating that frequencies are higher at later timesteps; compare e.g. "giraffe" and "alligator" in Figure 11 for examples of high mean and mode and low



(a) Sarah (Brown corpus)

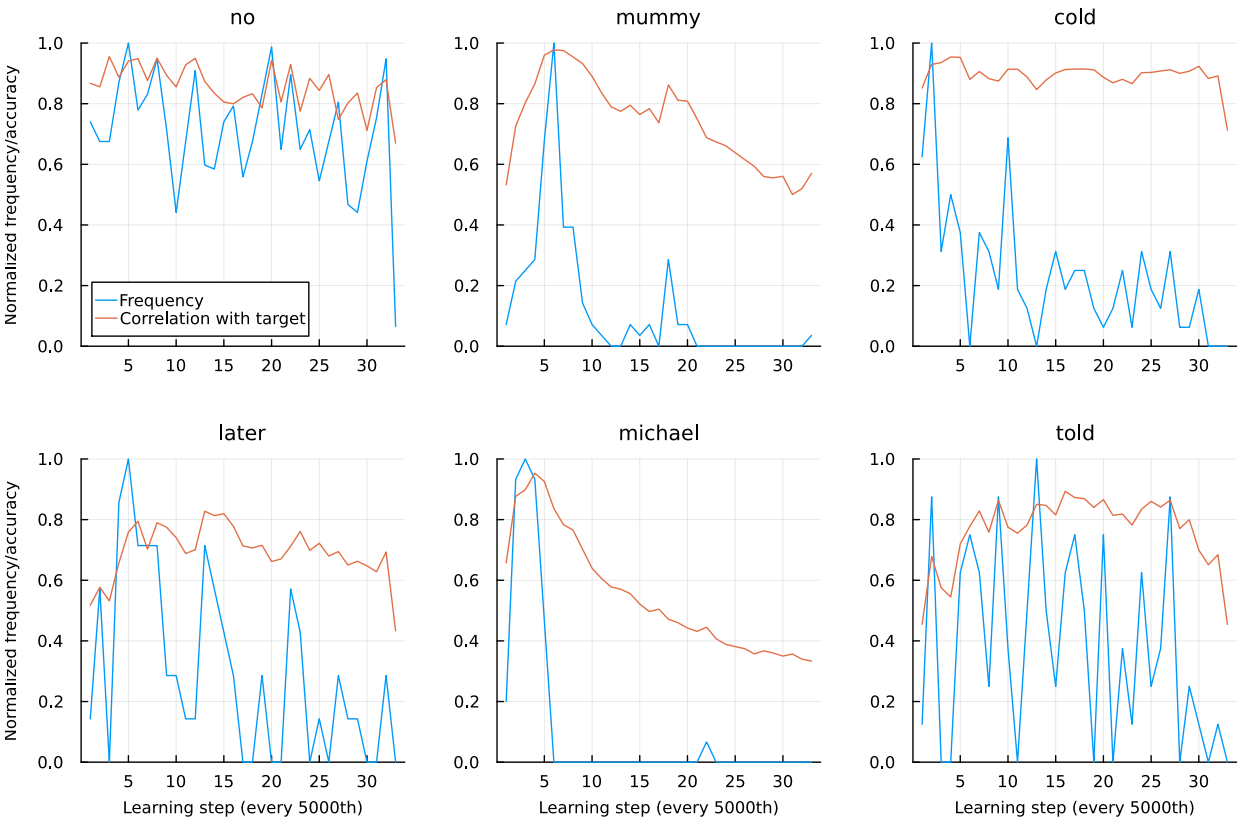


(b) Lily (Providence corpus)

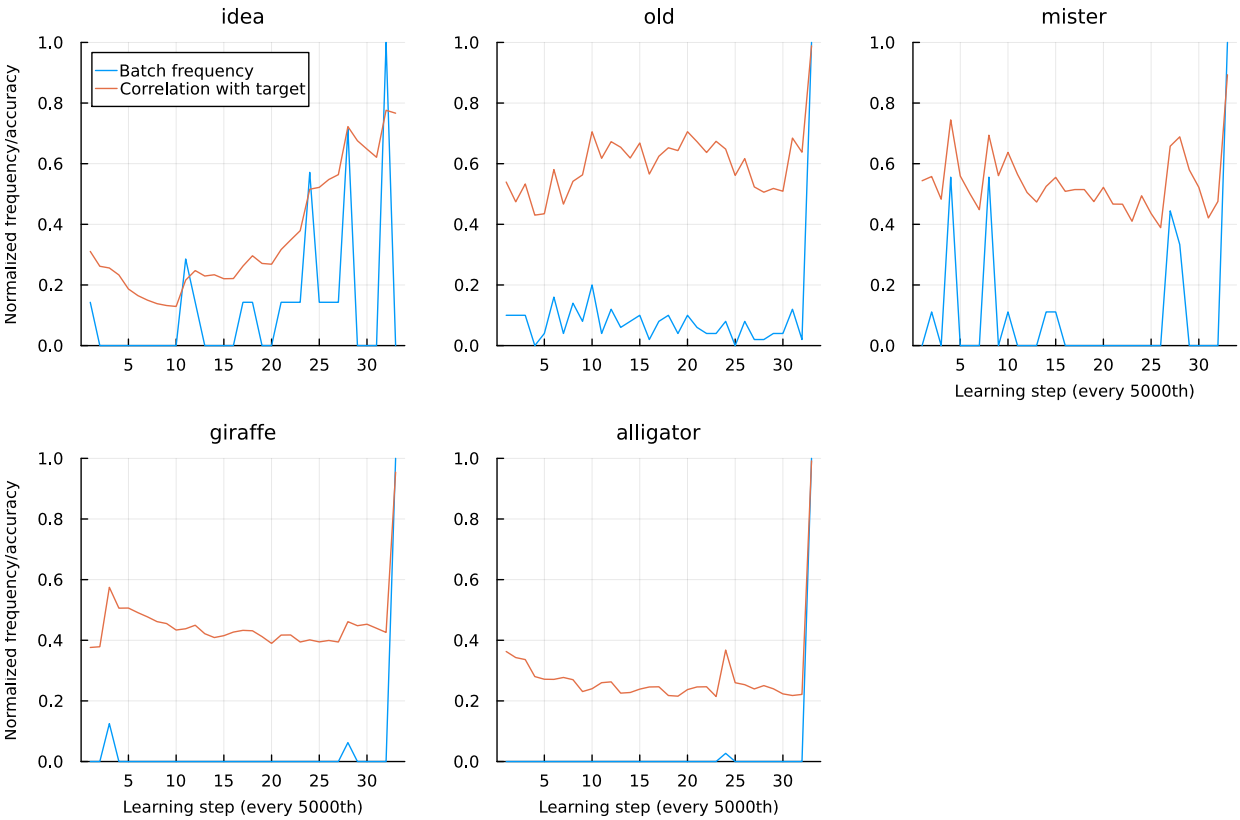
Figure 10: Correlation of WHL learned predicted semantics with their targets against correlations of FIL learned predicted semantics with their targets, for different learning rates.

skewness, with "mummy" and "michael" in Figure 11 for examples of low mean and mode and high skewness), the better WHL performed than FIL. To illustrate, we show the effect of the mean on performance in the upper row of Figure 12. Additionally, when kurtosis was entered (differentiated by whether skewness was positive or negative), an interesting effect emerged: for negative skewness values (high frequency at later time steps), more positive kurtosis (peakier distribution) yielded an advantage of WHL, while a peakier distribution of positive skewness values lead to an advantage of FIL (see bottom row of Figure 12). All these results confirmed our qualitative analysis above.

To summarise, some interesting phenomena related to order information get lost when using FIL. For higher learning rates, WHL overall yields higher accuracies, possibly because it is sensitive to the burstiness of frequency during the course of learning. On the other hand, for words that are equally distributed across learning events, the predictions of WHL and FIL are very similar, and the two methods thus result in highly similar correlations with their respective target semantics (see also Milin et al., 2020).



(a) FIL over WHL



(b) WHL over FIL

Figure 11: Individual tokens where FIL and WHL correlations to target differ clearly, taken from "Sarah", trained with WHL and $\eta = 0.01$. (a) shows cases where FIL outperforms WHL, (b) where WHL outperforms FIL. In (a), frequency tends to decrease over time, while in (b) the opposite is the case. Frequencies are normalised by their maximal frequency inside a learning batch of 5000 learning events.

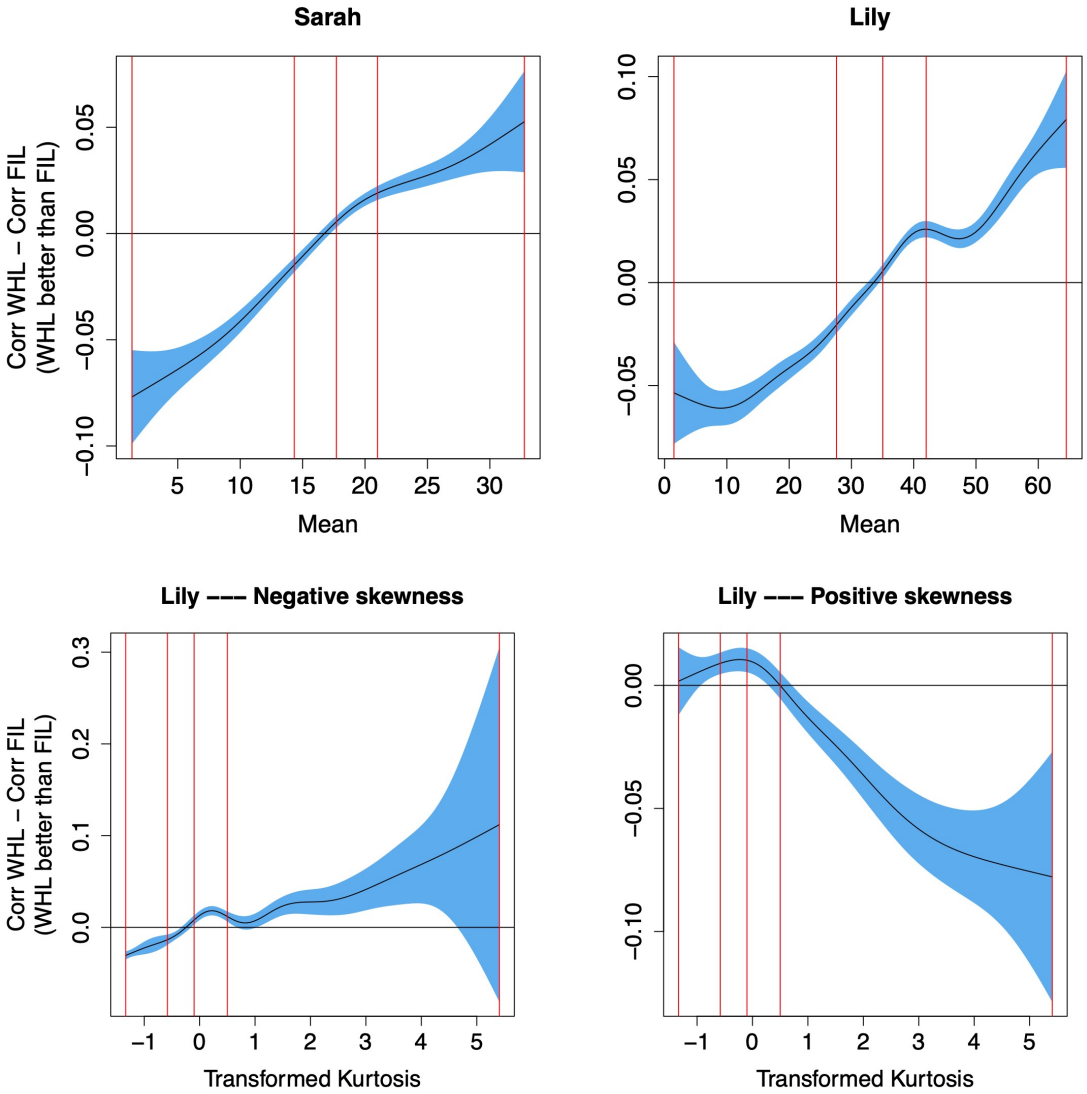


Figure 12: Predicting the difference between target correlations in WHL and in FIL from frequency distribution of words across time. **Top row:** The higher the mean (i.e. higher frequencies at later time steps, see "giraffe" and "alligator" in the lower row of Figure 11 for examples of words with a high mean), the better WHL performs than FIL. The vertical red lines represent the 0%, 25%, 50%, 75%, and 100% percentiles. **Bottom row:** For negative skew (higher frequencies at later time steps), the peakier the distribution (higher kurtosis), the larger the advantage of WHL over FIL, and vice versa for positive skew. Kurtosis was transformed to reduce outlier effects, details in Supplementary Materials.

9 Discussion

We have introduced a new way for estimating mappings between form and meaning in the Discriminative Lexicon Model (DLM; Baayen et al., 2018a, 2019) that takes frequency of use into account, Frequency-Informed Learning (FIL), complementing incremental learning with the learning rule of Widrow-Hoff (WHL) and endstate learning using multivariate multiple regression (EL). Each of these methods has advantages as well as disadvantages.

9.1 Three methods for computing mappings in the DLM

WHL enables trial to trial learning and hence is, in principle, an excellent choice for datasets with learning events that are ordered in time. Examples of such datasets are child-directed speech in the CHILDES database ordered by the age of the addressee and the time-series of reaction times in mega-experiments (Heitmeier et al., 2022). The disadvantage of WHL is that it is computationally demanding, and prohibitively so for large datasets.

FIL offers a computationally lean way of taking frequency of use into account, but it is insensitive to the order of learning events. It is therefore an excellent choice for datasets with learning events that are unordered, which is typically the case for data compiled from corpora or databases. For large datasets with temporally ordered learning events, FIL can be applied to a sequence of datasets with increasing sample sizes to probe how learning develops over time. How exactly such sequential modeling compares with WHL is a topic for further research.

Models using EL are not computationally demanding, but they are also not sensitive to the frequencies with which learning events occur. For usage-based approaches to language (see, e.g., Bybee, 2010), this is a serious drawback. Nevertheless, EL has an important advantage of its own: it provides a window on what can be learned in principle, with infinite experience. In other words, EL is a good analytical tool for any datasets for which a type-based analysis is appropriate or insightful. For instance, if one's interest is in how well Dutch final devoicing can be mastered on the basis of subliminal learning only, the EL model informs us that a comprehension accuracy of 83% can be reached (see also Heitmeier et al., 2023). When measures are gleaned from an EL model and used as predictors for aspects of lexical processing, it will typically be necessary to include a measure of frequency of use as a covariate.

With FIL, however, we have an analytical instrument that integrates experience into mappings between form and meaning. It obviates the practical necessity, when using WHL, of scaling frequencies down, nor is a log-transform of usage required. The latter is particularly undesirable as it artefactually boosts performance for low-frequency words while degrading performance for high-frequency words.

An open question when it comes to comparing the three methods is whether measures of the models' accuracy should be type-based, i.e. all words contribute equally to the overall accuracy measure, or whether it should be token-based. Previous models have usually been evaluated based on type-accuracy, but a high token-accuracy in word recognition might be of more practical use in every day life than a high type- but low token-accuracy, where a model is able to recognise many low frequency words but struggles with many of the high frequency words with which speakers are constantly surrounded.

9.2 The relationship between word frequency and lexical decision reaction times

A surprising property of FIL is that the correlation r of the predicted semantic vector with its gold standard target semantic vector emerges as a key to understanding two findings in lexical decision tasks. We found that FIL is crucial for modelling a stronger priming effect of segmental information compared to tone information in an auditory lexical decision task in Mandarin Chinese. For unprimed lexical decisions as available in the Dutch Lexicon Project, reasoning that greater correlation should afford shorter decision times, we used $1 - r$ as an approximation of simulated decision times. We found that $1 - r$ based on FIL provided much improved prediction accuracy compared to $1 - r$ based on EL. Moreover, FIL also provides insights into the non-linear nature

of the word frequency effect in the lexical decision task. We showed that the mirror-sigmoid relation between empirical decision times and log frequency emerging from a GAM analysis also characterizes the functional relation between 'simulated decision times' $1 - r$ and log frequency. This suggests that FIL successfully filters usage through discriminative learning to obtain estimates of how well the meanings of words are understood.

This finding fits in well with a recent debate that was re-sparked by Murray and Forster (2004) arguing that rank-transformed frequencies account for lexical decision reaction times better than log-transformed ones and that, therefore, serial-search models should not be discounted as models of word recognition.[8] Recently, Kapatsinski (2022) showed that log frequencies transformed by the logistic function (a function frequently used in deep learning models) predict reaction times in the same way as a rank-transformation, implying that the linear relationship between rank frequency and reaction times is not necessarily evidence in favour of the serial search model. Since the DLM is not a classification model, we do not make use of the logistic function but use correlation to compute how close a word's predicted meaning is to its true meaning. The estimated functional relation between -1000/RT and $1 - r$ estimated with FIL is close to linear (see Figure 7, panel (2,1)), and a very similar partial effect emerges when the untransformed RTs are regressed on $1 - r$. Thus, a linear relationship between a FIL-based predictor, $1 - r$ (or equivalently, r) and reaction times falls out directly of the DLM, without requiring further transformations. This provides further evidence in favour of theories suggesting that frequency effects arise due to the distributional properties of words forms and meanings during learning.

A question that we leave to future research is whether measures derived from FIL mappings will obviate the need to include frequency of use as a covariate, reduce the variable importance of this predictor, or complement it. One possible complication here is that while frequency-informed mappings make the DLM predictions more realistic, they naturally also create a confound with word frequency, which needs to be teased apart from other (e.g. semantic) effects captured by the DLM (see Gahl and Baayen, 2023, for further discussion).

What complicates such comparisons is that frequencies based on corpora typically concern lexical use in a language community, whereas individual speakers have very different experiences with lower-frequency words, depending on their interests, professions, education, and reading/speaking habits (e.g. Kuperman and Van Dyke, 2013; Baayen et al., 2016). For instance, about 5% of the Dutch words in the present dataset are unknown to the fourth author. These are low-frequency words that the EL gets correct in 30 out of 35 cases. By contrast, the FIL gets only 2 out of these 35 cases correct. In a similar vein, Diependaele et al. (2012) showed that in the DLP, when reacting twice to the same lowest-frequency word stimuli in the DLP they only agree in about 50% of cases, bringing their accuracy to chance level.

The inter- and intra-variability of subjects' word knowledge might also be a reason why FIL underestimates the variance of reaction times in the Dutch Lexicon Project for low-frequency words. Naturally, FIL, being based on 'community' frequencies, is not able to account for this effect. Our results, however, highlight the importance of modelling not only the mean but also the variance of reaction times, as it might provide a further window into speakers' word recognition process and differentiating between different models attempting to account for the observed behavioural data.

When evaluating the accuracy of a FIL model, having the closest proximity to the semantic vector of the target word is probably too stringent a criterion. Especially for lower-frequency words, having a rank among the top k (for low k) nearest neighbors may be a more reasonable criterion. The reason is that for lower frequency words, knowledge of what words may mean is only approximate. For instance, although the names of gemstones such as jasper, ruby, tourmaline, and beryl will be known to many readers, picking out the ruby from four reddish pieces of rock requires more precise knowledge than is available to the authors. Especially for lexical decision making,

[8] However, note that the variance of word frequencies is similar in magnitude to the frequency itself (under the assumption that word frequencies are Poisson-distributed). By moving from frequencies to ranks, differences in frequency that seem large but that will vary substantially across samples are reduced to much smaller differences in ranks. In the light of these considerations, the excellent predictivity of rank for reaction times is due to distributional properties of the language in combination with sampling error, rather than due to serial searches in frequency ordered mental word lists.

being 'close enough' may be good enough for a yes-response (see Heitmeier et al., 2022, for detailed discussion of lexical decision making).

An additional complication is that low-frequency words often have multiple, equally rare, meanings. (By contrast, for high-frequency words, one often finds that of a set of possible meanings, one is dominant.) By way of example, the low-frequency Dutch word *bras* in our dataset can denote 'uncooked but peeled rice', 'junk', and 'a specific set of ropes used on sailing ships'. This may provide further explanation of why the correlation measure $(1 - r)$ underestimates the variance in decision latencies for lower-frequency words as compared to higher-frequency words.

9.3 Computational modelling of word frequency effects

Our findings are particularly enlightening when comparing them to accounts of word frequency effects in previous models of word recognition. To answer the three key questions regarding the word frequency effects identified in the introduction, in the DLM, frequency effects arise as a consequence of the input distribution it is trained on. Similarly to other network models, the effect of word frequency is stored in the model's mapping weights. Since the DLM is an extremely simple computational model involving only linear mappings, a fast closed-form solution is available, which, as demonstrated, is very similar to training the model incrementally on unordered data (as is done in all deep learning models utilising backpropagation, such as the triangle model Seidenberg and McClelland, 1989). As such, it enables the modeller to answer many questions related to the word frequency effect, such as that of individual differences, in a computationally very lean way. Finally, we demonstrate that the DLM is able to account for the non-linear relationship between frequency and reaction times in lexical decision without requiring to use a log-transformation (cf. McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982; Seidenberg and McClelland, 1989). Moreover, the DLM can not only account for frequency effects but also enables investigation of the influence of words' more finegrained characteristics on reaction times. Crucially, these are not independent. The DLM's predicted reaction times also depend on words' form and meaning properties, not only their frequency distribution.

To put the present results in a broader perspective, consider the characterization given by Breiman et al. (2001) of machine learning on the one hand and statistics on the other hand. They argue that machine learning aims at obtaining accurate predictions. How these predictions are obtained and why a technique works is not of interest. By contrast, statistics aims to formulate a model that could have generated the data. Within this characterisation — which may be too extreme (e.g. Shmueli, 2010) — FIL is much closer to statistical modeling than to machine learning, and it is surprising to us how much can be achieved simply with a form of weighted multivariate multiple regression. We hope that FIL will prove to be a useful tool not only for modeling data with 'community' frequencies but also for exploring, by means of simulation, what the consequences are of individual differences in usage for lexical processing.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645, and by the European Research Council, project WIDE-742545. The authors thank members of the Quantitative Linguistics group for helpful feedback and discussions, and Yu-Hsiang Tseng for helpful comments on an earlier version of this manuscript.

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.

Baayen, R. H. (2001). *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

Baayen, R. H. (2005). Data mining at the intersection of psychology and linguistics. In Cutler, A., editor, *Twenty-first century psycholinguistics: Four cornerstones*, pages 69–83. Erlbaum, Hillsdale, New Jersey.

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, 5(3):436–461.

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

Baayen, R. H., Chuang, Y.-Y., and Heitmeier, M. (2018b). WpmWithLdl: Implementation of word and paradigm morphology with linear discriminative learning. R package Version 1.2.20.

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., Dijkstra, T., and Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language*, 36:94–117.

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.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Brown, R. (1973). A first language: The early stages. *Cambridge, MA: Harvard University.*

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., and Böhl, A. (2011). The word frequency effect. *Experimental psychology.*

Brysbaert, M., Mandera, P., and Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, 27(1):45–50.

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.

Bybee, J. (2010). *Language, usage and cognition.* Cambridge University Press, Cambridge.

Bybee, J. L. and Hopper, P. J. (2001). *Frequency and the Emergence of Linguistic Structure*, volume 45. John Benjamins Publishing.

Chuang, Y.-Y., Bell, M. J., Banke, I., and Baayen, R. H. (2021). Bilingual and multilingual mental lexicon: a modeling study with linear discriminative learning. *Language Learning*, 71(S1):219–292.

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

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

Demuth, K., Culbertson, J., and Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and speech*, 49(2):137–173.

Denistia, K. and Baayen, R. H. (2021). The morphology of indonesian: Data and quantitative modeling. *The Routledge Handbook of Asian Linguistics. Routledge.*

Diependaele, K., Brysbaert, M., and Neri, P. (2012). How noisy is lexical decision? *Frontiers in psychology*, 3:348.

Ernestus, M. T. C. and Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in dutch. *Language*, 79(1):5–38.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., and Pallier, C. (2010). The french lexicon project: Lexical decision data for 38,840 french words and 38,840 pseudowords. *Behavior research methods*, 42:488–496.

Fon, J. and Hsu, H.-j. (2007). Positional and phonotactic effects on the realization of dipping tones in Taiwan Mandarin. In Gussenhoven, C. and Riad, T., editors, *Phonology and Phonetics, Tones and Tunes: Vol. 2. Experimental Studies in Word and Sentence Prosody*, pages 239–269. Mouton de Gruyter, Berlin.

Forster, K. (1976). Accessing the mental lexicon. In Wales, F. and Walker, E., editors, *New approaches to language mechanisms*, pages 257–286. Amsterdam: North-Holland.

Forster, K. I. (1979). Levels of processing and the structure of the language processor. *Sentence processing*, pages 27–85.

Forster, K. I. (1994). Computational modeling and elementary process analysis in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 20(6):1292.

Gahl, S. and Baayen, R. H. (2023). Time and thyme again: Connecting spoken word duration to models of the mental lexicon. *Under revision for Language.*

Grave, É., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

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.

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, H. (2022). How trial-to-trial learning shapes mappings in the mental lexicon: Modelling lexical decision with linear discriminative learning. *arXiv preprint arXiv:2207.00430*.

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

Heitmeier, M., Chuang, Y.-Y., Luo, X., and Baayen, H. (2023). Linear Discriminative Learning: Theory and implementation in the julia package JudiLing. *Manuscript, University of Tübingen*.

Ho, A. T. (1976). The acoustic variation of Mandarin tones. *Phonetica*, 33(5):353–367.

Howie, J. M. (1974). On the domain of tone in mandarin. *Phonetica*, 30(3):129–148.

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.

Kapatsinski, V. (2022). The logistic perceptron accounts for rank frequency effects in lexical processing. In Nixon, J., Tomaschek, F., and Baayen, R. H., editors, *Proceedings of the Second International Conference on Error-Driven Learning in Language (EDLL 2022)*, pages 16–17.

Keuleers, E., Diependaele, K., and Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 dutch mono-and disyllabic words and nonwords. *Frontiers in psychology*, 1:174.

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.

Kuperman, V. and Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, 39(3):802.

Landauer, T., Foltz, P., and Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Lee, C.-Y. (2007). Does horse activate mother? processing lexical tone in form priming. *Language and Speech*, 50(1):101–123.

Luo, X. (2021). JudiLing: An implementation for Linear Discriminative Learning in JudiLing (unpublished Master's thesis).

MacWhinney, B. (2000). The CHILDES project. *Tools for Analyzing Talk. Part*, 1.

McClelland, J. L. and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological review*, 88(5):375.

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

Milin, P., Madabushi, H. T., Croucher, M., and Divjak, D. (2020). Keeping it simple: Implementation and performance of the proto-principle of adaptation and learning in the language sciences. *arXiv preprint arXiv:2003.03813*.

Miwa, K. and Baayen, H. (2021). Nonlinearities in bilingual visual word recognition: An introduction to generalized additive modeling. *Bilingualism: Language and Cognition*, 24(5):825–832.

Morton, J. (1969). Interaction of information in word recognition. *Psychological review*, 76(2):165.

Morton, J. (1979a). Facilitation in word recognition: Experiments causing change in the logogen model. *Processing of visible language*, pages 259–268.

Morton, J. (1979b). Word recognition. *Psycholinguistics: Series 2. Structures and processes*, pages 107–156.

Murray, W. S. and Forster, K. I. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, 111(3):721.

Nieder, J., Chuang, Y.-Y., van de Vijver, R., and Baayen, R. (2021). Comprehension, production and processing of maltese plurals in the discriminative lexicon. *PsyArXiv 10.31234/osf.io/rkath*.

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.

Nusbaum, H. C. (1985). A stochastic account of the relationship between lexical density and word frequency. Technical report, Indiana University. Research on Speech Perception, Progress Report #11.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of experimental psychology*, 81(2):275.

Rescorla, R. A. (1967). Pavlovian conditioning and its proper control procedures. *Psychological review*, 74(1):71.

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.

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.

Rubenstein, H., Lewis, S. S., and Rubenstein, M. A. (1971). Homographic entries in the internal lexicon: Effects of systematicity and relative frequency of meanings. *Journal of verbal learning and verbal behavior*, 10(1):57–62.

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.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., and Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods*, 51(4):1928–1941.

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.

Seidenberg, M. S. and McClelland, J. L. (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.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, pages 289–310.

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.

Sun, C. C., Hendrix, P., Ma, J., and Baayen, R. H. (2018). Chinese lexical database (cld) a large-scale lexical database for simplified mandarin chinese. *Behavior Research Methods*, 50:2606–2629.

Tomaschek, F., Tucker, B., and Baayen, R. H. (2018). Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistic Vanguard*, 4:1–13.

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

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.

Xu, Y. and Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech communication*, 33(4):319–337.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., and Matsumoto, Y. (2020). Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30. Association for Computational Linguistics.