

Is deeper always better? Replacing linear mappings with deep learning networks in the Discriminative Lexicon Model

Maria Heitmeier, Valeria Schmidt,
Hendrik P.A. Lensch and R. Harald Baayen

Recently, deep learning models have increasingly been used in cognitive modelling of language. This study asks whether deep learning can help us to better understand the learning problem that needs to be solved by speakers, above and beyond linear methods. We utilise the Discriminative Lexicon Model introduced by Baayen and colleagues, which models comprehension and production with mappings between numeric form and meaning vectors. While so far, these mappings have been linear (Linear Discriminative Learning, LDL), in the present study we replace them with deep dense neural networks (Deep Discriminative Learning, DDL). We find that DDL affords more accurate mappings for large and diverse datasets from English and Dutch, but not necessarily for Estonian and Taiwan Mandarin. DDL outperforms LDL in particular for words with pseudo-morphological structure such as *chol+er*. Applied to average reaction times, we find that DDL is outperformed by frequency-informed linear mappings (FIL). However, DDL trained in a frequency-informed way (‘frequency-informed’ deep learning, FIDDL) substantially outperforms FIL. Finally, while linear mappings can very effectively be updated from trial-to-trial to model incremental lexical learning, deep mappings cannot do so as effectively. At present, both linear and deep mappings are informative for understanding language.

1 Introduction

Over the last decade, deep learning models¹ have been remarkably successful, culminating in the most recent large language models (e.g. [OpenAI, 2023](#); [Anthropic, 2023](#); [Gemini Team, 2023](#)). Increasingly, deep learning has also made its way into cognitive modelling of language with studies exploring whether deep learning architectures can model aspects of language processing. Modelling efforts have been targeted at roughly four groups of models: predicting how humans form known and novel wordforms, such as inflections (e.g. [Rumelhart and](#)

¹Throughout this paper, we use the term ‘deep learning’ to refer to neural networks with at least one hidden layer (with a non-linear activation function). While it can be argued whether a neural network with a single hidden layer can indeed be called ‘deep’, the definitions of deep learning models usually only require such models to have “multiple (non-linear) processing layers” (e.g. [Arnold et al., 2011](#); [LeCun et al., 2015](#)) and for reasons of simplicity we therefore include models with a single non-linear hidden layer in the definition. Crucially, and contrary to models without any hidden layers and no activation functions, such models are able to learn non-linear functions (e.g. [Webb and Lowe, 1990](#); [Widrow and Lehr, 1990](#)).

McClelland, 1986; Cardillo et al., 2018; Kirov and Cotterell, 2018; Corkery et al., 2019; McCurdy et al., 2020), using activations in deep learning models to directly predict behavioural data or brain signals (e.g. Bhatia and Richie, 2022; Salewski et al., 2023; Huang et al., 2024; Caucheteux and King, 2022; Schrimpf et al., 2021; Hale et al., 2022; Goldstein et al., 2022), using deep learning components in theoretical models of lexical processing (e.g. Seidenberg and McClelland, 1989; Harm and Seidenberg, 2004; Magnuson et al., 2020; Scharenborg, 2008), and finally, leveraging the iterative learning ability of deep learning models to model language acquisition (e.g. Plunkett et al., 1992; Li et al., 2007).

While these efforts have often been quite successful, using deep learning comes at the cost of ease of interpretation (but see e.g. Cammarata et al., 2020; Elhage et al., 2021; Bricken et al., 2023; Linzen and Baroni, 2021). Deep learning models are universal function approximators, implying that if powerful enough, they can approximate any continuous function with arbitrary accuracy (Hornik et al., 1989). For cognitive modelling, however, pure accuracy is not the end-goal: the question is whether increased precision afforded by deep learning models actually helps with predicting (cognitive) behaviour.

In the present paper, we address this question for the Discriminative Lexicon Model (DLM, Baayen et al., 2018, 2019), a model of (single) word comprehension and production. The model has been used to model the morphology of a wide range of languages (e.g. Chuang et al., 2020; Nieder et al., 2023; Heitmeier et al., 2023a; Chuang et al., 2023; van de Vijver and Uwambayinema, 2022) but has also been successful at predicting behavioral data such as reaction times and acoustic durations (e.g. Gahl and Baayen, 2024; Stein and Plag, 2021; Heitmeier et al., 2023b; Cassani et al., 2019; Chuang et al., 2021; Baayen and Smolka, 2020; Saito, 2023). Word forms are represented by high-dimensional binary vectors and semantics by high-dimensional real-valued vectors. In the DLM’s original formulation, comprehension is modelled with a linear mapping from word forms to word meanings and vice versa for production. Linear mappings are highly interpretable and surprisingly powerful, but also inherently limited by their linearity. In the present study we use the properties of linear and deep learning to investigate whether computational modeling can help us to better understand the learning problem that needs to be solved by speakers. Since we investigate single-word learning rather than word sequences and we believe that in the context of cognitive modelling the analysability of models is very important, we focus on using simple dense neural networks rather than more complex deep learning architectures such as Recurrent Neural Networks (RNNs) or transformers. We are specifically interested in whether, and under what circumstances, deep learning generates insights above and beyond what more easily interpretable, linear models can tell us.

This paper is structured as follows: Section 2 briefly describes the DLM, the dense neural networks (termed Deep Discriminative Learning, DDL) we used to replace the linear mappings and how they were trained. Section 3 shows that DDL generally affords higher mapping accuracy, particularly on the training data, and explores for which words DDL outperforms LDL. Section 4 investigates DDL’s ability to predict lexical decision reaction times, both averaged across individuals and on a trial-by-trial basis. Finally, Section 5 discusses the results and concludes the paper.

2 Methods

2.1 Data

In the present study we used four datasets²:

British Lexicon Project (BLP): Large-scale British English visual lexical decision database

²Data, code and analysis scripts can be found in the Supplementary Materials at <https://osf.io/tyfdr/>.

(Keuleers et al., 2012). The final dataset contained 28,482 word forms. We modelled this dataset with trigrams and word embeddings grounded in vision from Shahmohammadi et al. (2021).

Dutch Lexicon Project (DLP): Large-scale Dutch visual lexical decision database (Keuleers et al., 2010). The final dataset contained 13,675 word forms. We modelled this dataset with trigrams and fasttext embeddings (Grave et al., 2018).

Estonian: This dataset was first introduced in Chuang et al. (2020). The final dataset contained 4,093 wordforms of 230 lemmas, thus on average 17.8 wordforms per lemma (range from 2 to 28). We modelled this dataset with trigrams and fasttext embeddings (Grave et al., 2018).

Taiwan Mandarin: This dataset was first introduced in Chuang and Baayen (2021) based on Fon (2004). It contained 7,349 words with one to five syllables. We used a phone-based representation where vowels are combined with a number indicating lexical tone (e.g. 一下子, eng. ‘abruptly’, is represented as i1.x.ia4.z.i15) and fasttext embeddings (Grave et al., 2018).

The four datasets pose different challenges: English and Dutch have limited inflectional morphology, therefore the challenge of the BLP and DLP datasets lies in modelling a large number of different lemmas. The dataset for Estonian, with its rich inflectional system, challenges our models to accurately map many different forms for only a few different lemmas. Finally, the challenge of the Mandarin dataset lies in modelling both segmental (phone) and suprasegmental (tone) information.

2.2 Models

In the Discriminative Lexicon Model (DLM) comprehension is modelled as a mapping from a word’s form to its meaning, and production as a mapping from meaning to form. The simplest way to represent wordforms is by first splitting wordforms into letter/phone n-grams and then marking the presence and absence of n-grams in high-dimensional binary vectors. The form vectors of all wordforms in the lexicon are stored in a form matrix \mathbf{C} . Word meanings are likewise represented by high-dimensional vectors, usually taken from pre-trained word embedding spaces such as word2vec or fasttext (Mikolov et al., 2013; Grave et al., 2018). Details on which embedding vectors were used for which dataset can be found in Section 2.1. The embedding vectors are then likewise stored in a semantic matrix \mathbf{S} . In Linear Discriminative Learning (LDL), comprehension is then modelled via a linear mapping \mathbf{F} such that

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

and production via \mathbf{G}

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

with $\hat{\mathbf{S}}$ and $\hat{\mathbf{C}}$ indicating predicted \mathbf{S} and \mathbf{C} matrices respectively, borrowing notation from statistics. There are several methods available for calculating \mathbf{F} and \mathbf{G} , such as standard least squares regression which optimises across all word forms equally, or “frequency-informed learning” which uses a form of weighted regression to inform the mappings about word forms’ frequencies (Heitmeier et al., 2023a). We used various methods throughout this study and give more detail in the relevant sections.

Once \mathbf{F} and \mathbf{G} are in place, they can be used to predict semantic and form vectors for seen or unseen word forms w via

$$\hat{\mathbf{s}}_w = \mathbf{c}_w \mathbf{F}$$

$$\hat{\mathbf{c}}_w = \mathbf{s}_w \mathbf{G}$$

where \mathbf{c}_w and \mathbf{s}_w are the form and meaning vectors of w and $\hat{\mathbf{s}}_w$ and $\hat{\mathbf{c}}_w$ the model’s predicted vectors. A thorough introduction to LDL can be found in Heitmeier et al. (2025a).

In Deep Discriminative Learning (DDL) we replace \mathbf{F} and \mathbf{G} with dense neural networks³:

$$\begin{aligned}\hat{\mathbf{S}} &= f(\mathbf{C}) \\ \hat{\mathbf{C}} &= g(\mathbf{S}).\end{aligned}$$

Since multi-layer neural networks can in principle approximate any mathematical function (with some limitations, Hornik et al., 1989), we denote them with f and g .

Please refer to Figure 1 for an overview of the number of hidden layers and dimensions for each model in this paper (one model per dataset and comprehension/production). We used Rectified Linear Units (ReLUs)⁴ as activation function throughout. The comprehension models were trained with a mean squared error loss. For production models, we opted to regard the prediction of which n-grams are part of a wordform as a multi-class multi-label classification problem where all n-grams are “classes” and multiple n-grams can be selected as “labels”. As commonly done in multi-class multi-label classification models, we therefore added a sigmoid function after the final layer of the production models to restrict the final predicted values between 0 and 1 (approximating probabilities) and trained them with a binary cross-entropy loss. Note that adding a sigmoid + binary cross-entropy loss to the model in this way does not add any additional trainable parameters. We used a batchsize of 512, except for the Estonian dataset where we used a batchsize of 256. Unless indicated otherwise, we used the Adam optimizer (Kingma, 2014) with a learning rate of 0.001. Further justification for these modelling choices can be found in Heitmeier et al. (2025b). Other training details depend on the experiment and are laid out in the sections below.

We note here that we do not regard the different architectures across the different datasets and tasks as being informative about the cognition of speakers of the respective languages. Instead, we think that the architectures are determined by the datasets. The datasets vary considerably in size which influences the networks’ required sizes. The languages’ structures presumably also influence the architectures, but even regarding this point, we are wary of making any claims about speakers’ cognition given that the kind of neural networks we are using are not particularly biologically plausible and how words were selected varies considerably across datasets. Instead, we view the presented architectures as the optimal models for solving the specific mapping problems for the specific datasets (within the scope of hyperparameters tested). As such, we are interested in what can theoretically be learned by a statistical system and testing whether what is learned by such a system allows insights about behaviour as well.

3 Does DDL improve mapping accuracy?

First, we need to verify that DDL does indeed improve accuracy of the mappings compared to LDL. Here, we focus both on performance on seen data as well as unseen data. We split the data into an approximate 80/10/10 split for training/validation/test data. The data was split in such a way that all letter n-grams in the validation and test data also occurred in the training data. We trained models for all four languages for a maximum of 500 epochs, with

³Note that we also experimented with LSTMs (Hochreiter and Schmidhuber, 1997) and transformer architectures (Vaswani et al., 2017) but ultimately found that for the problem at hand simple dense neural networks provided the best generalisation accuracy while also being directly comparable to the linear setup.

⁴Due to space limitations, we refrain here from defining terms commonly used in machine learning literature. A more accessible introduction to Deep Discriminative Learning including term definitions can be found in Chapter 11 of Heitmeier et al. (2025a).

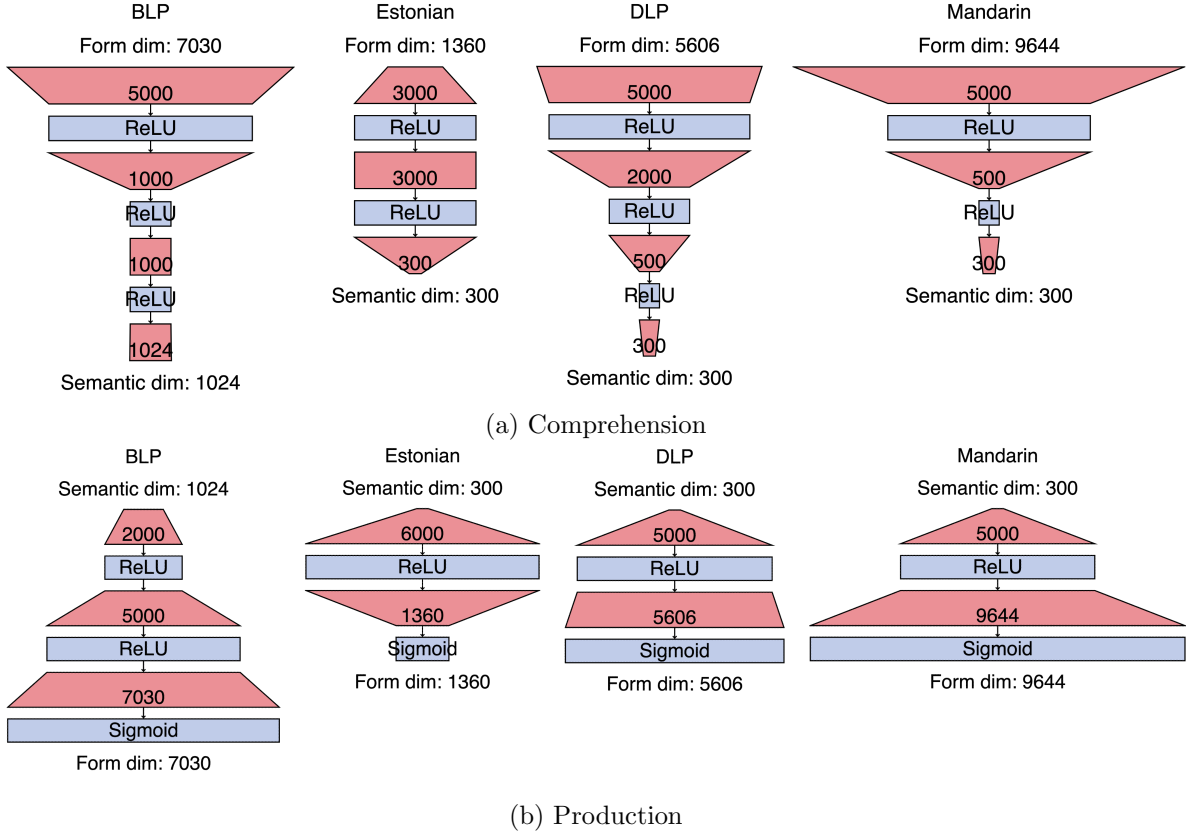


Figure 1: DDL model architectures. The red boxes represent dense layers (matrix multiplication plus bias) with the output dimension printed inside, blue boxes represent non-linearities. Box widths reflect the input and output dimensions of the layers. Figure from Heitmeier et al. (2025b).

early stopping when validation accuracy did not increase for 20 epochs for the comprehension and 50 epochs for the production models. As a baseline, we also trained LDL models on the joint training and validation data (since LDL does not require a validation set)⁵. We evaluated all models using correlation accuracy: if the predicted vector was more correlated with its target vector than with all vectors in the remaining data, it was counted as correct. We also computed accuracy@10 for the best models for each datasets. For accuracy@10, a mapping is counted as correct, if the target vector is among the 10 most correlated vectors. The results can be found in Figure 2.

It is clear that DDL outperforms LDL for comprehension only for the English data where the challenge lies in mapping a diverse set of meanings. For Estonian and Mandarin on the other hand, LDL already performs surprisingly well and DDL only improves on LDL for production. However, DDL outperforms LDL for all datasets in production and the general increase in training accuracy across all four languages indicates that DDL does indeed afford higher mapping precision.

Next, we investigated for which words DDL outperforms LDL. We hypothesised that DDL would perform better than LDL for words which are harder to distinguish from other words.

⁵Even though DDL is not directly trained on the validation data, it is nevertheless used in training to select the epoch after which to stop training when performance on the validation data is best. Thus, even though the validation data is not directly used to calculate the loss and update the model’s weights, it nevertheless influences the trained model. Due to this crucial role in the *training* of the DDL models, it is important to train the linear models on both the training and validation data.

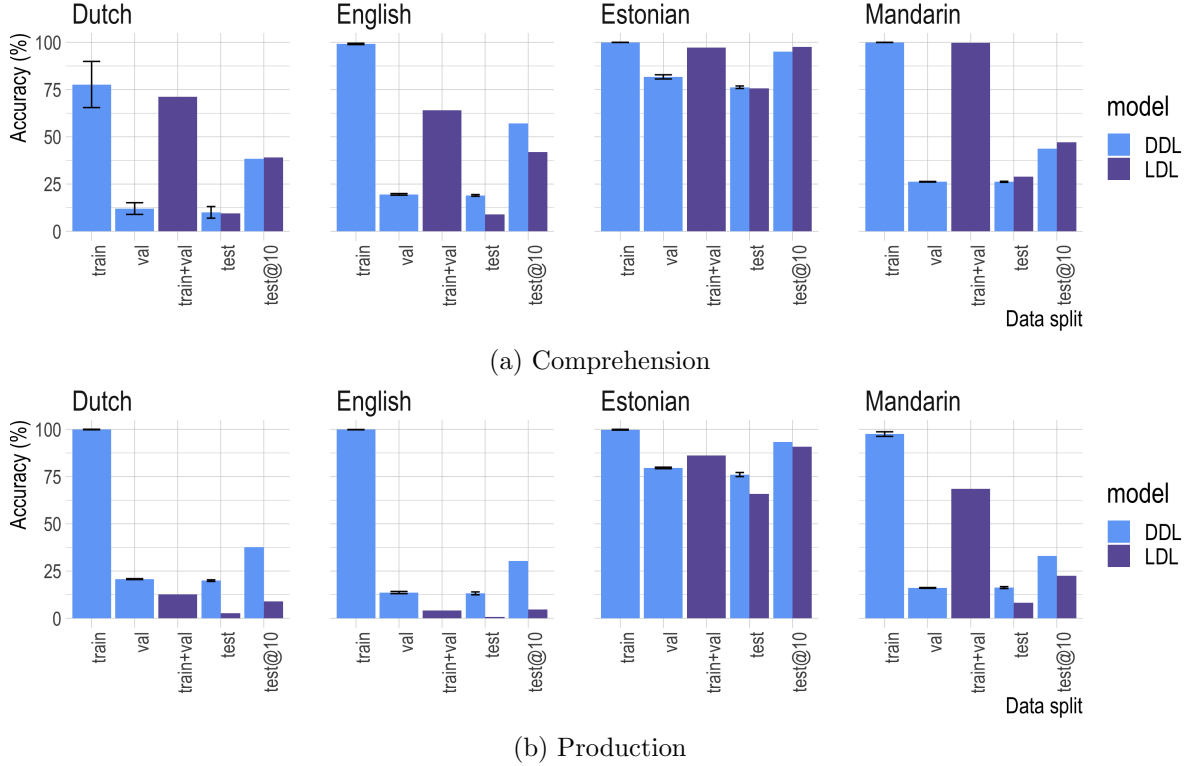


Figure 2: Correlation accuracy for DDL and LDL models for a British English, Dutch, Estonian and Taiwan Mandarin dataset. For DDL models, accuracies are averaged across 10 models, the error bar shows the standard deviation. DDL clearly outperforms LDL for Dutch and English, and less clearly for Estonian and Mandarin. Test@10 indicates the accuracy@10 on the test data for the best model.

These are words with n-grams that are shared with many other words, shorter words and words in a dense semantic neighbourhood. We used these predictors (plus word frequency as a control) to predict the difference in correlation between predicted and target vectors in LDL and DDL using a Generalised Additive Model (GAM, Wood, 2011; more details on how we operationalised these concepts can be found in Heitmeier et al., 2025b). Figure 3 shows that DDL indeed outperforms LDL for words with higher cue overlap (i.e. a word’s cues occur in many other words), shorter words and words in a denser orthographic neighbourhood (higher Coltheart’s N, Coltheart, 1977). Moreover, it performs better on words with higher frequency. While semantic density (measured as the correlation between the target semantic vector and its closest semantic neighbour) seems to have a U-shaped effect, this is due to collinearity with the other predictors (Friedman and Wall, 2005): in a GAM with only semantic density as a predictor, the measure shows the expected effect of an advantage for DDL for words in denser semantic neighbourhood.

We also qualitatively examined words that showed particularly stark differences in correlation, favoring either LDL or DDL. The results can be found in Table 1. LDL outperforms DDL if words contain unique cues such as *px* or *oz*. DDL on the other hand shows a clear advantage for words which contain suffixes expressing a range of different meanings such as *barriers* or *choler* which all contain trigrams which are often used in comparatives. Consequently, LDL places them in the semantic neighbourhood of other comparatives, while DDL maps them to their correct neighbourhood. This ability of DDL to work around the re-use of trigrams to express different meanings points to an interesting issue: language is full of re-use of sub-lexical

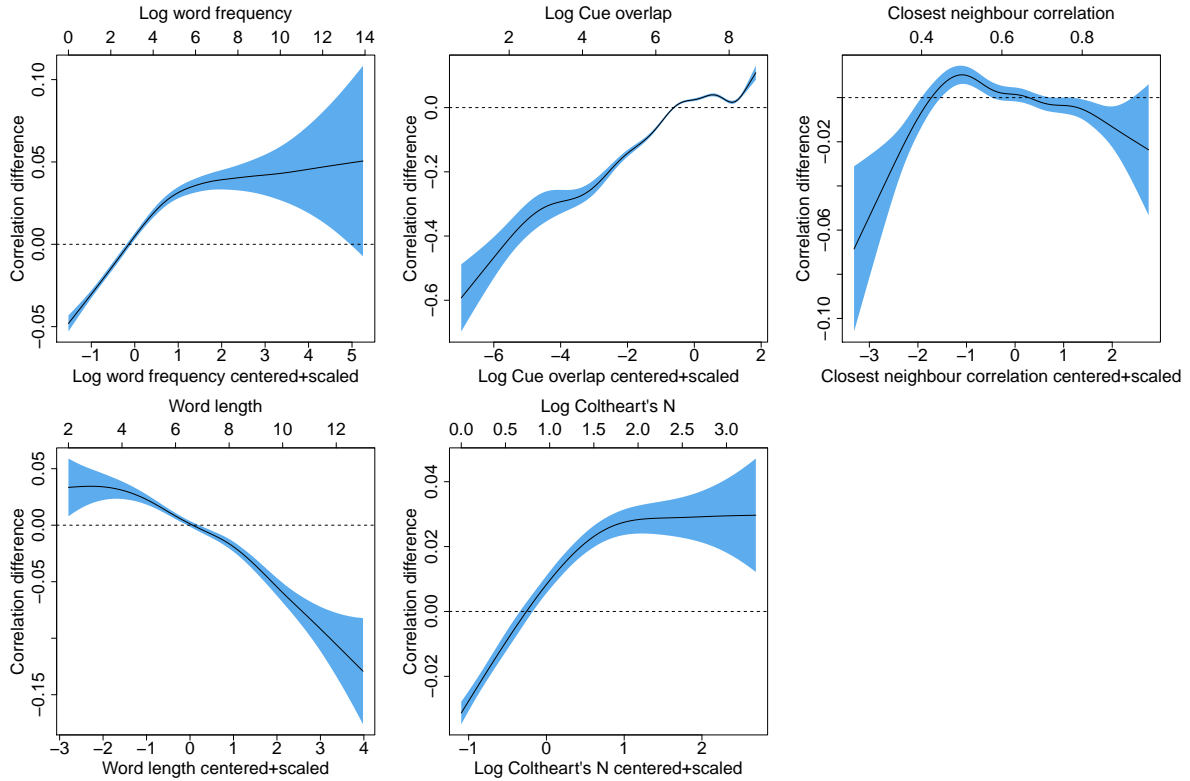


Figure 3: Measures of between-word similarity predicting the difference between target correlation in DDL and LDL (target correlation DDL - target correlation LDL). For higher values of correlation difference, DDL outperforms LDL. The lower x-axes show the centered and scaled values as entered into the GAMs; the upper x-axes show the original values for reference. We find that in general, DDL outperforms LDL for words that are similar to other words (i.e. words with higher cue overlap, a dense orthographic neighbourhood, shorter words and more frequent words). Figure adapted from [Heitmeier et al. \(2025b\)](#).

features, such as the *-s* ending, which can express plurality (see also [Shafaei-Bajestan et al., 2022](#); [Plag et al., 2015](#)), third person singular, or possessive or simply be part of a word’s stem, or the *-er* ending, which can express varied meanings such as comparative (e.g., *hungrier*), agent (e.g., *teacher*), instrument (e.g., *poker*), or something associated with an action (e.g., *diner*) or can denote the origin of a person (e.g., *Londoner*) ([Plag, 2003](#), p.89). Any system modelling morphology needs to be able to deal with this “bricolage” ([Lévi-Strauss, 1962](#)). Apparently, DDL is much more effective at doing so than LDL. This does not seem to be just an issue of precision: both Table 1 and the accuracies@10 in Figure 2 suggest that for LDL, the correct target word does not even make it to the top 10 most correlated words. A replication of this analysis for Dutch can be found in the Supplementary Materials.

4 Does DDL improve prediction of behavioural data?

Having ascertained that DDL does indeed generally improve mapping precision, particularly for the BLP and DLP, we were interested next in whether this also means that DDL outperforms LDL at predicting behavioural data. To this end we used visual lexical decision data from the BLP and DLP. In the following, we present results of predicting reaction times averaged across participants and predicting reaction times on a trial-by-trial basis following

Word	Semantic neighbours LDL	Semantic neighbours DDL	Correlation difference
by	by, standby, bye, written, thereby, brigades, dismissed, copied, outplayed, routed	bypassed, policed, weeded, cleared, entrenched, enforced, watered, encroached, deterred, imposed	-0.51
px	px, width, height, border, size, square, margin, padding, serif, background	px, margin, apex, equals, border, serif, font, zero, cornered, width	-0.50
my	my, her, me, your, own, myself, his, friend, our, dearest	my, friend, dearest, dear, proudest, thankful, hearts, beloved, fathers, coolest	-0.44
if	if, unless, sure, else, whether, because, assume, not, does, should	if, unless, assume, sure, whereas, therefore, neither, likely, maybe, suppose	-0.40
oz	oz, ounce, ounces, bottle, quart, gallon, jar, servings, cologne, spray	oz, ounce, ounces, bottle, juice, gallon, shampoo, tumbler, cleanser, pint	-0.38

(a) LDL outperforms DDL

Word	Semantic neighbours LDL	Semantic neighbours DDL	Correlation difference
lungs	lunches, lunge, slogs, dragged, luncheons, lunges, lurches, clamoured, jaunts, dives	lungs, lung, chest, breath, windpipe, coughing, exhale, larynx, breathe, throat	0.54
scarves	carves, carving, starves, carvers, carver, corpses, gouges, carved, sews, sculpts	scarves, scarfs, sweaters, stoles, blankets, hats, headbands, afghans, robes, capes	0.55
leans	leaches, cleanses, scoundrels, deign, nazis, cleans, traitors, betters, purveys, kikes	leans, pulls, mutters, nods, pushes, straightens, crouches, grabs, squeezes, smirks	0.56
barriers	livelier, chattier, sturdier, flashier, rowdier, trendier, sketchier, ruder, heartier, fleshier	barriers, barrier, impede, roadblocks, blocking, curbs, hurdle, shackles, hinder, doorways	0.56
choler	chunkier, wiser, moister, heartier, blunter, fattier, gunsmith, braver, plainer, trimmer	choler, slackness, rancour, wherefore, quinsy, foulness, rapine, glumness, stoutness, largeness	0.59

(b) DDL outperforms LDL

Table 1: Words for which LDL outperforms DDL the most (a) and for which DDL outperforms LDL the most (b). The two “semantic neighbours” columns show the words of the 10 semantic vectors which are most correlated with the vectors predicted by LDL and DDL respectively. The rightmost column shows the difference in target correlation between LDL and DDL. LDL outperforms DDL for words with unique n-grams, while DDL outperforms LDL for words with endings which can express various meanings (e.g. the *-er* ending). In the latter case, LDL tends to place words in the semantic neighbour of words with the same ending, while DDL puts them in the neighbourhood of words with similar meanings.

[Heitmeier et al. \(2023b\)](#).

4.1 Predicting average reaction times

For both the DLP and BLP we trained five comprehension models:

- Endstate-of-Learning (EL): LDL model optimised equally across all words in the DLP/BLP.
- Frequency-informed learning (FIL): LDL model trained in a frequency-informed way using a form of weighted regression ([Heitmeier et al., 2023a](#)).
- Deep Discriminative Learning (DDL): DDL model trained on the 90% most frequent words in the DLP/BLP. Training was stopped after accuracy on the remaining 10% of words did not improve for 20 epochs.
- Endstate-DDL (EDDL): DDL model trained on all words in the DLP/BLP for 2000 epochs to model a theoretical endstate of learning where all words are learned optimally.

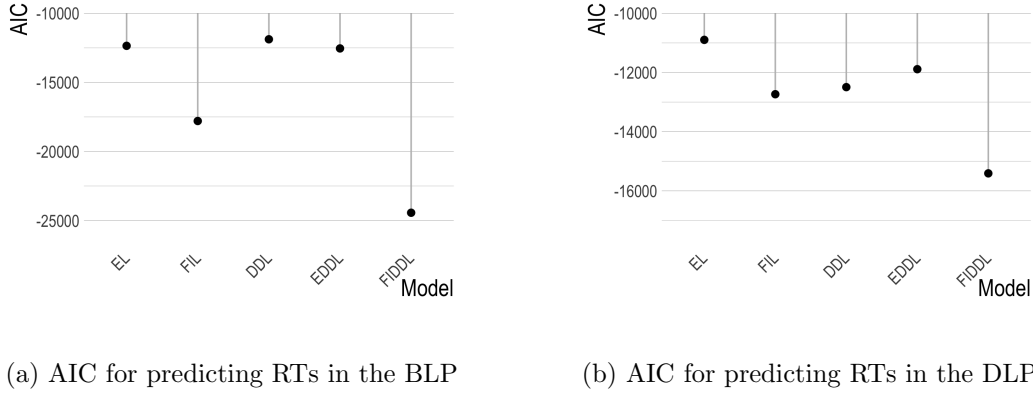


Figure 4: Target correlation taken from an EL, FIL, DDL, EDDL and FIDDL model predicting reaction times in the BLP and DLP. Figure adapted from [Heitmeier et al. \(2025b\)](#).

- Frequency-informed DDL (FIDDL): DDL model trained on the full token distribution of words in the DLP/BLP, i.e. the model was trained on each word as many times as its frequency count.

We used CELEX ([Baayen et al., 1995](#)) frequencies for both the BLP and DLP throughout. After training, we obtained the correlations between the predicted and the target semantic vectors for each model (“target correlation”). Reaction times were transformed as follows:

$$\text{RTinv} = \frac{-1000}{\text{RT}}$$

We used the target correlation measure to predict RTinv in one GAM per discriminative learning model. Target correlation was a highly significant predictor for reaction times across all models and both datasets ($p < 0.0001$; summary tables for the GAMs can be found in the Supplementary Materials). Figure 4 shows the fit of the five models in terms of the Akaike Information Criterion (AIC, [Akaike, 1998](#))⁶ for both the DLP and BLP. First, we see that for both datasets, frequency-informed methods (FIL, FIDDL) outperform their non-frequency-informed counterparts (EL, DDL, EDDL). Regarding the difference between DDL and LDL, DDL and EDDL outperform EL for the DLP, but not for the BLP. Interestingly, while the correlation accuracy for EDDL was clearly higher than for DDL (DLP: 100% for EDDL compared to 93% for DDL; BLP: 100% compared to 89%), when it comes to predicting reaction times, EDDL only shows a slight advantage over DDL for the BLP and performs worse for the DLP. This indicates that better mapping accuracy does not necessarily entail better performance at predicting behavioural data. FIDDL outperforms all other models. From this we can conclude that while DDL does not necessarily outperform LDL, FIDDL clearly outperforms LDL both without and with frequency information. We will return to both these points in the final discussion.

4.2 Predicting trial-to-trial reaction times

Next, we turn to using DDL models for predicting more fine-grained data, namely lexical decision data at the per-participant trial-to-trial level. [Heitmeier et al. \(2023b\)](#) hypothesised that humans’ lexical knowledge is continuously changing and that this continuous change can be modelled with the DLM. They simulated the lexical decision experiment reported in the

⁶AIC measures model fit while punishing model complexity. Lower AIC values indicate better model fit.

BLP. For each participant they ran one simulation where they updated the LDL mappings in the DLM using the Widrow-Hoff (Widrow and Hoff, 1960) learning rule (dynamic simulation) and one where they did not update the mappings (static simulation). After applying one learning step of the Widrow-Hoff learning rule, the error between a predicted and target vector is slightly lower than before the update. This allows updating the mappings after each encountered stimulus in the lexical decision experiment in our simulations (details in Heitmeier et al., 2023b, 2025b). Measures from these simulations were then used to predict individual reaction times in the BLP. They found that for a majority of subjects, dynamic simulations provided a better model fit to reaction times, particularly for words (as opposed to nonwords).

Here, we asked whether we can replicate these findings with DDL. To this end we repeated the analysis presented in Heitmeier et al. (2023b) for the first 39 subjects in the BLP, replacing LDL mappings with DDL models and applying one backpropagation step (instead of the Widrow-Hoff learning rule) for updating the models after each trial. As in the linear simulations, each backpropagation step included calculating the loss of only this trial’s predicted vector and using this loss to update the model’s weights. Details on the simulation procedure, the measures used for predicting reaction times and the statistical modelling can be found in Heitmeier et al. (2023b) and Heitmeier et al. (2025b).

We found that the AIC of all DDL-based models was worse than that of the LDL-based counterparts for all participants, for both words and nonwords and static/dynamic simulations (average AIC difference was -108.5/-128.4 AIC points for words for static/dynamic simulations and -128.3/-118.1 AIC points for nonwords for static/dynamic simulations). While for LDL we found that 72% of dynamic simulations showed a better fit than static simulations for words (33% for nonwords), for DDL 74% of dynamic simulations outperformed static ones for words (62% for nonwords). Figure 5 gives an overview over the average difference in AIC as well as the percentage of participants showing an advantage of trial-to-trial learning.

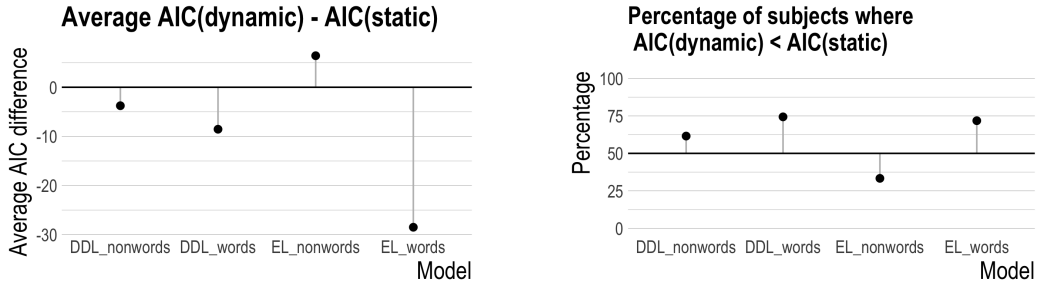


Figure 5: Comparison between static and dynamic simulations using DDL and LDL for predicting per-participant reaction times in the BLP. For words, the effect of dynamic simulations compared to static ones is larger for LDL, resulting in a bigger difference in AIC. Figure adapted from Heitmeier et al. (2025b).

When plotting the difference between the AICs of static and dynamic simulations with DDL against that with LDL (Figure 6), we found that they correlate for words, that is, DDL and LDL generally agreed on which subjects show an effect of trial-to-trial learning as modelled by the dynamic simulation. No such correlation was present for nonwords.

We conclude from this analysis that DDL is not able to model fine-grained trial-to-trial effects as well as LDL. For words, DDL shows an advantage of learning for a similar number of subjects as LDL but the difference between dynamic and static is smaller for DDL than

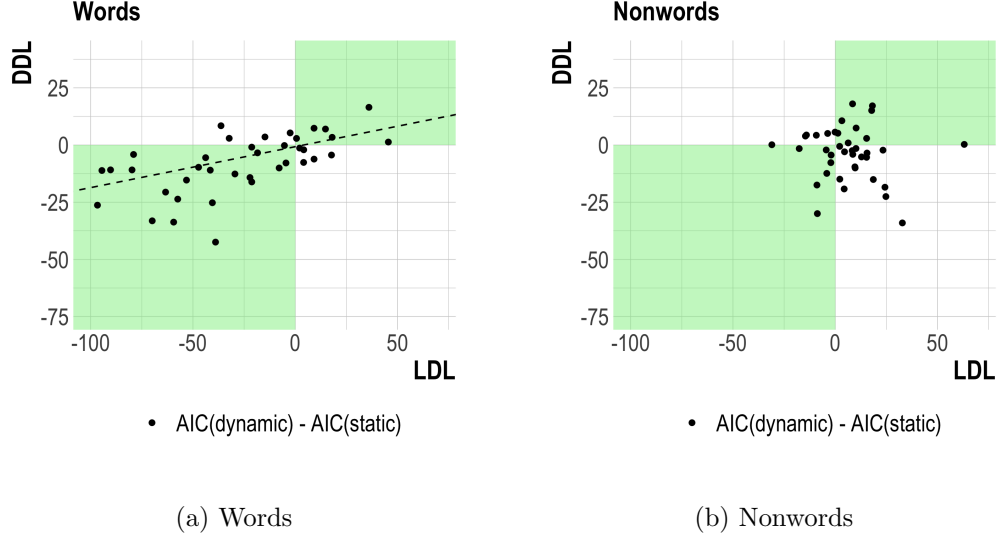


Figure 6: Comparison of AIC differences between dynamic and static simulations for DDL and LDL. Each dot represents one participant. The x-axis denotes the difference in AIC between dynamic and static simulations for the LDL models and the y-axis for the DDL models. Negative difference values indicate that the dynamic simulation outperforms the static simulation in terms of AIC, while positive difference values indicate the opposite. If a dot falls within a green shaded area, LDL and DDL agree on whether dynamic simulations provide a better model fit than static simulations for the respective participant. For words, the AIC differences are correlated ($r = .48, p = 0.001$); for nonwords no correlation is present ($p = 0.4715$). Figure adapted from [Heitmeier et al. \(2025b\)](#).

for LDL. At least for words, DDL and LDL largely agree on which subjects show an effect of trial-to-trial learning. A possible reason for the reduced effect of trial-to-trial learning in DDL is that a more accurate initial mapping may lead to less pronounced error feedback in the update step, thus each learning step having a smaller impact on subsequent trials. Moreover, the larger number of weights in the DDL networks might also mean that weight updates are distributed more evenly across weights and not centered as strongly on the weights of cues shared with subsequent words.

5 Discussion and Conclusion

In this paper we explored whether replacing the linear mappings (Linear Discriminative Learning, LDL) in the Discriminative Lexicon Model (DLM) with dense neural networks (termed Deep Discriminative Learning, DDL) improves the model’s mapping accuracy. More crucially for cognitive modelling however, we also investigated whether such improved mapping accuracy also increases the model’s performance at predicting behavioural data. We found that across four languages (English, Dutch, Taiwan Mandarin, Estonian) DDL always improved training accuracy and also accuracy on held-out data for English on comprehension, the dataset with the largest number of different lemmas. For production, DDL performed better than LDL across both training and test data for all four datasets. When predicting average lexical decision reaction times, we found that a frequency-informed variant of DDL (FIDDL) where the model is trained on each word according to its frequency outperformed all other

methods (including frequency-informed LDL, FIL). However, when predicting fine-grained trial-to-trial lexical decision reaction times, DDL performs worse than LDL, even though the two generally agree as to which subjects show a better model fit when based on simulations where the DLM is dynamically updated after each trial.

It is worthwhile to focus on a few aspects of these results in more detail. First, we find a notable difference across datasets: In comprehension, DDL only reliably improves mapping accuracy over that of LDL for English, but not for Mandarin and Estonian. This may to some extent be explainable by the respective data sizes: the English dataset is larger (28,000 word forms) than the Estonian and Mandarin ones (6,000 and 3,000). However, the nature and diversity of the datasets may also play role. The English and Dutch datasets have a much larger number of different lemmas with few inflected word forms per lemma, while the Estonian dataset only has 230 lemmas with a large number of (relatively regular) inflected forms. For Mandarin, tone information contributes to a much larger number of cues (9,644, compare this to the much larger English dataset, which only has 7,030 cues), and as we saw in Section 3, fewer shared cues increase the performance of LDL models in comprehension. This suggests that the more semantically diverse a dataset, the more it profits from modelling with DDL. In production, on the other hand, all four models profit from DDL. In Heitmeier et al. (2025b) we found that this can be accounted for to a significant extent by the loss function: we used the binary cross-entropy loss to train a linear model with the same number of trainable parameters as LDL but with an added sigmoid function and found that it achieved almost the same accuracy as the deep learning models. However, the advantage of DDL over LDL in the present setup may also be due to the representations used. Chuang et al. (2024) found that LDL performs surprisingly well at mapping contextualised embeddings to pitch contours, and preliminary results suggest that in this setup, DDL does not improve over LDL.⁷ This indicates that when both input and output are sufficiently discriminable, LDL suffices.

Second, notice that while DDL models outperformed LDL models in terms of pure mapping accuracy at least for the English dataset, they do not necessarily outperform LDL at predicting reaction times: DDL outperforms LDL at predicting average reaction times in the Dutch Lexicon Project (DLP), but not in the British Lexicon Project (BLP). The same is true for modelling lexical decision reaction times at the individual subject-level: even leaving aside trial-to-trial learning, the DDL models never perform better than the LDL models for static simulations. This highlights that pure mapping accuracy is not enough for predicting reaction times. However, it also leads to a further point, namely that modern machine learning and artificial intelligence may in some aspects be too good: human speakers make a lot of mistakes and are not particularly accurate. For example, participants in the DLP agree with themselves at chance level for low frequency words (below 10 per million) when tested on the same words twice (Diependaele et al., 2012). In the study of Ramscar et al. (2013), young adults (mean age 21.1 years) performed at chance level for low frequency words in an English lexical decision task (Balota et al., 1999). Moreover, the question arises of whether fully optimized performance in the DLM would be in any way helpful for predicting reaction times. Dense neural networks can in principle learn any continuous function to an arbitrary point of accuracy when given enough parameters (Hornik et al., 1989). In the case of the DLM this would mean that all predicted semantic and form vectors would be entirely accurate, thus losing any information we can gain from the DLM above and beyond simply using binary form vectors and pre-trained embeddings. The DLM’s imprecisions and mistakes offer valuable insights into possible processing bottlenecks.

Third, examining the words that DDL maps more accurately than LDL is illuminating for linguistic analysis. We found that LDL outperforms DDL for words which contain very unique

⁷However, mapping accuracy from form (low-dimensional) to embeddings (high-dimensional) did improve compared to linear mappings, which is perhaps unsurprising given the difference in dimensionality.

trigrams (e.g. *#px*, *#oz*), while DDL outperforms LDL when words contain trigrams which express a range of different functions such as *-er*. The re-use of building blocks is ubiquitous in language. While humans are able to deal with this “bricolage” relatively effortlessly, it is nevertheless a processing bottleneck which LDL struggles with. In this way, the comparison between LDL and DDL shows which components of single word processing may be particularly challenging. This finding could also have implications for theories of morphological priming (e.g. Taft and Forster, 1975; Rastle and Davis, 2008): our results suggest that LDL may place pseudomorphological words (e.g. *barriers*, *choler*) in the semantic neighbourhood of morphologically complex words (e.g. *livelier*, *chunkier*). This could explain why morphologically transparent and morphologically opaque words are primed in a similar way (e.g. Creemers et al., 2020) even without morphological decomposition (for linear DLM models capturing morphological priming effects, see Baayen and Smolka, 2020; Chuang et al., 2023). Again, however, this result may, at least in part, be due to the representations used. An examination of the final segment of words ending in *-er* in the Buckeye corpus (Pitt et al., 2007) reveals that they have more varied pronunciations in nouns than in adjectives. These results dovetail well with the hypothesis that the nouns, which have much more varied meanings, have more differentiated realizations of the exponent — a distinction that is neutralized in the orthography. This in turn suggests that deep learning is required for *-er* words in English because the representations are impoverished, which opens the possibility that a linear mapping based on actual articulation might suffice.

Fourth, our results highlight the importance of the data distribution the models are trained on for cognitive modelling. Using a training/validation/test split is not a plausible mechanism for training cognitive models: after all, we usually encounter high frequency (seen) words, with the occasional low frequency (possibly unseen) word interspersed in-between. FIDDL (frequency-informed DDL) therefore has a much more plausible learning setup: it is trained on the full token distribution. By calculating the model’s accuracy across all words in the dataset we get a more realistic picture of how the model would fare with a real-world distribution of tokens (see also Heitmeier et al., 2023a, for a discussion of token-based accuracy calculations). Surprisingly, this accuracy is very high (97% and 87% for Dutch and English respectively) even though the model saw 40% of the words in the English and 9% of the words in the Dutch dataset only 10 times or fewer. Such a setup also quells concerns about the performance of computational models on held-out data (Pinker and Ullman, 2002). Humans are able to generalise to unseen data, but in reality, word distributions are skewed towards high frequency words. Evidently, both humans and FIDDL perform excellently in such a setting. Nonetheless, one might argue that particularly in languages with rich morphology such as Finnish, speakers may still need to generalise to low frequency unseen word forms regularly. As shown by Nikolaev et al. (2024) for a set of some 50,000 Finnish inflected nouns, EL can already perform very well on both training data and held-out low-frequency words. To address this question with further precision, a natural next step is to evaluate performance of FIDDL on these and similar data sets. In any case, these results once again highlight the importance of training models of lexical processing in a frequency-informed way when one is interested in a usage-based analysis of the data. The frequency-informed FIDDL and FIL models by far outperform their non-frequency informed counterparts when predicting average reaction times in the BLP and DLP.

To summarise, researchers interested in modelling a language system or behavioural data with the DLM may well profit from using DDL when they are working with a particularly large, diverse dataset which LDL struggles with and/or when they require a higher mapping accuracy for their modelling task. However, they should be aware that using DDL may not necessarily improve the model’s ability to predict behavioural data (unless trained in a frequency-informed way) and that DDL is costly: it requires experimenting with a number of

hyperparameters as well as significantly more computational resources and training times.

Where does this leave us with regards to our initial question about deep learning’s role in understanding the learning problem which speakers face? Our comparison between LDL and DDL in the DLM lead us to a few conclusions. While deep learning models are generally more accurate, they are not a “tried-and-true” method for increasing cognitive modelling performance: pure accuracy improvements do not necessarily afford better performance at predicting behavioural data and may also not improve our understanding of the underlying processes. However, investigating cases where deep learning models clearly outperform linear models can nevertheless point to processing bottlenecks in a purely linear system which can be solved at the expense of additional parameters in deep learning models.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC number 2064/1 – Project number 3907276455, and by the European Research Council, project WIDE-742545. The authors thank Yu-Ying Chuang, Tino Sering and Louis ten Bosch for helpful discussions.

References

- Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer.
- Anthropic (2023). Model Card and Evaluations of Claude Models. *Accessed on Nov 21, 2023*.
- Arnold, L., Rebecchi, S., Chevallier, S., and Paugam-Moisy, H. (2011). An introduction to deep learning. In *European symposium on artificial neural networks (ESANN)*.
- Baayen, R. H., Chuang, Y.-Y., and Heitmeier, M. (2018). 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., 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.
- Balota, D. A., Cortese, M. J., and Piloti, M. (1999). Item-level analyses of lexical decision performance: Results from a mega-study. In *40th annual meeting of the Psychonomics Society*, page 44.
- Bhatia, S. and Richie, R. (2022). Transformer networks of human conceptual knowledge. *Psychological review*, 131(1):271–306.
- Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Aspell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume,

- T., Carter, S., Henighan, T., and Olah, C. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>, accessed on April 28, 2024.
- Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., Schubert, L., Voss, C., Egan, B., and Lim, S. K. (2020). Thread: Circuits. *Distill*. <https://distill.pub/2020/circuits>, accessed on April 27, 2024.
- Cardillo, F. A., Ferro, M., Marzi, C., and Pirrelli, V. (2018). Deep learning of inflection and the cell-filling problem. *Italian Journal of Computational Linguistics*, 4(1):57–75.
- Cassani, G., Chuang, Y.-Y., and Baayen, R. H. (2019). On the semantics of non-words and their lexical category. *Journal of Experimental Psychology: Learning, Memory, Cognition*, 46(4):621–637.
- Caucheteux, C. and King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(135).
- Chuang, Y.-Y. and Baayen, R. H. (2021). Discriminative learning and the lexicon: NDL and LDL. In *Oxford research encyclopedia of linguistics*. Oxford University Press Oxford.
- Chuang, Y.-Y., Bell, M. J., Tseng, Y.-H., and Baayen, R. H. (2024). Word-specific tonal realizations in mandarin. *arXiv preprint arXiv:2405.07006*.
- Chuang, Y.-Y., Kang, M., Luo, X., and Baayen, R. H. (2023). 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.
- Chuang, Y.-Y., Vollmer, M.-l., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., and Baayen, R. H. (2021). The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behaviour Research Methods*, 53:945–976.
- Coltheart, M. (1977). Access to the internal lexicon. *The psychology of reading*.
- Corkery, M., Matushevych, Y., and Goldwater, S. (2019). Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3868–3877.
- Creemers, A., Davies, A. G., Wilder, R. J., Tamminga, M., and Embick, D. (2020). Opacity, transparency, and morphological priming: A study of prefixed verbs in dutch. *Journal of Memory and Language*, 110:104055.
- Diependaele, K., Brysbaert, M., and Neri, P. (2012). How noisy is lexical decision? *Frontiers in psychology*, 3.
- Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T.,

- Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2021/framework/index.html>, accessed on April 28, 2024.
- Fon, J. (2004). A preliminary construction of Taiwan Southern Min spontaneous speech corpus. Technical report, Tech. Rep. NSC-92-2411-H-003-050, National Science Council, Taiwan.
- Friedman, L. and Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, 59(2):127–136.
- Gahl, S. and Baayen, R. H. (2024). Time and thyme again: Connecting english spoken word duration to models of the mental lexicon. *Language*, 100(4):623–670.
- Gemini Team (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380.
- 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)*, pages 3483–3487.
- Hale, J. T., Campanelli, L., Li, J., Bhattasali, S., Pallier, C., and Brennan, J. R. (2022). Neurocomputational models of language processing. *Annual Review of Linguistics*, 8:427–446.
- Harm, M. W. and Seidenberg, M. S. (2004). Computing the meanings of words in reading: co-operative division of labor between visual and phonological processes. *Psychological review*, 111(3):662–720.
- Heitmeier, M., Chuang, Y., Axen, S., and Baayen, R. H. (2023a). Frequency-informed linear discriminative learning. *Frontiers in Human Neuroscience*, 17.
- Heitmeier, M., Chuang, Y.-Y., and Baayen, H. (2025a). *The Discriminative Lexicon: Theory, implementation in the Julia package JudiLing, and applications*. Manuscript, University of Tübingen.
- Heitmeier, M., Chuang, Y.-Y., and Baayen, R. H. (2023b). How trial-to-trial learning shapes mappings in the mental lexicon: Modelling lexical decision with linear discriminative learning. *Cognitive Psychology*, 146:101598.
- Heitmeier, M., Schmidt, V., Lensch, H., and Baayen, R. H. (2025b). Study 4: Deep discriminative learning: replacing linear mappings with deep learning networks in the Discriminative Lexicon Model. In *Mappings in the Discriminative Lexicon Model (PhD thesis)*. University of Tübingen.
- Hochreiter, S. and Schmidhuber, J. (1997). Long Short-term Memory. *Neural computation*, 9(8):1735–1780.
- Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

- Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., and Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.
- 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.
- Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kirov, C. and Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.
- Lévi-Strauss, C. (1962). *Savage mind*. University of Chicago.
- Li, P., Zhao, X., and Mac Whinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive science*, 31(4):581–612.
- Linzen, T. and Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.
- Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabi, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., et al. (2020). Earshot: A minimal neural network model of incremental human speech recognition. *Cognitive science*, 44(4):e12823.
- van de Vijver, R. and Uwambayinema, E. (2022). A word-based account of comprehension and production of Kinyarwanda nouns in the Discriminative Lexicon. *Linguistics Vanguard*, 8(1):197–207.
- McCurdy, K., Goldwater, S., and Lopez, A. (2020). Inflecting when there’s no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756. Association for Computational Linguistics.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Nieder, J., Chuang, Y.-Y., van de Vijver, R., and Baayen, H. (2023). A discriminative lexicon approach to word comprehension, production, and processing: Maltese plurals. *Language*, 99(2):242–274.
- Nikolaev, A., Baayen, R. H., and Chuang, Y.-Y. (2024). Analyzing finnish inflectional classes through discriminative lexicon models. Paper presented at the Digital Research Data and Human Sciences (DRDHum) conference, Joensuu, Finland, December 11, 2024.
- OpenAI (2023). GPT-4 Technical Report. *arXiv:2303.08774*.

- Pinker, S. and Ullman, M. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, 6(11):456–463.
- Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Folser-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release). *Columbus, OH: Department of Psychology, Ohio State University (Distributor)*.
- Plag, I. (2003). *Word-formation in English*. Cambridge University Press, Cambridge.
- Plag, I., Homann, J., and Kunter, G. (2015). Homophony and morphology: The acoustics of word-final s in english. *Journal of Linguistics*, pages 1–36.
- Plunkett, K., Sinha, C., Møller, M. F., and Strandsby, O. (1992). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, 4(3-4):293–312.
- Ramscar, M., Hendrix, P., Love, B., and Baayen, R. H. (2013). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon*, 8(3):450–481.
- Rastle, K. and Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language and cognitive processes*, 23(7-8):942–971.
- Rumelhart, D. and McClelland, J. (1986). On Learning the Past Tenses of English Verbs. In *Parallel Distributed Processing*, volume 2, page 216–271. MIT Press.
- Saito, M. (2023). *Enhancement effects of frequency: An explanation from the perspective of Discriminative Learning*. PhD thesis, University of Tübingen.
- Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., and Akata, Z. (2023). In-context impersonation reveals large language models’ strengths and biases. *Advances in Neural Information Processing Systems*, 36:1–14.
- Scharenborg, O. (2008). Modelling fine-phonetic detail in a computational model of word recognition. In *The 9th Annual Conference of the International Speech Communication Association*, pages 1473–1476. ISCA Archive.
- Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.
- Seidenberg, M. S. and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4):523–568.
- Shafaei-Bajestan, E., Moradipour-Tari, M., Uhrig, P., and Baayen, R. H. (2022). Semantic properties of english nominal pluralization: Insights from word embeddings. *arXiv preprint arXiv:2203.15424*.
- Shahmohammadi, H., Lensch, H. P., and Baayen, H. (2021). Learning zero-shot multifaceted visually grounded word embeddings via multi-task training. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 158–170.
- 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.

- Taft, M. and Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of verbal learning and verbal behavior*, 14(6):638–647.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Webb, A. R. and Lowe, D. (1990). The optimised internal representation of multilayer classifier networks performs nonlinear discriminant analysis. *Neural Networks*, 3(4):367–375.
- Widrow, B. and Hoff, M. (1960). Adaptive switching circuits. *1960 WESCON Convention Record Part IV*.
- Widrow, B. and Lehr, M. A. (1990). 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, 78(9):1415–1442.
- 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.