Running head: DENSITY AND DISTINCTIVENESS

Density and distinctiveness in early word learning: Evidence from neural network simulations

Samuel David Jones and Silke Brandt

Department of Linguistics and English Language

Lancaster University, Lancaster, UK

Keywords: *word learning; memory; phonology; neural network; connectionist modeling.*

Word count: 4911

Correspondence concerning this article should be addressed to Sam Jones, Department of Linguistics and English Language, County South, Lancaster University, Lancaster, UK, LA1 4YL. Email: sam.jones@lancs.ac.uk

An online repository containing all data and code required to reproduce this analysis can be found at: https://osf.io/2qk5j/

Abstract

High phonological neighborhood density has been associated with both advantages and disadvantages in early word learning. High density may support the formation and fine-tuning of new word sound memories; a process termed lexical configuration (e.g. Storkel, 2004). However, new high-density words are also more likely to be misunderstood as instances of known words, and may therefore fail to trigger the learning process (e.g. Swingley & Aslin, 2007). To examine these apparently contradictory effects, we trained an autoencoder neural network on 587,954 word tokens (5497 types; including mono- and multi-syllabic words of all grammatical classes) spoken by 279 caregivers to English-speaking children aged 18 to 24 months. We then simulated a communicative development inventory administration and compared network performance to that of 2292 children aged 18 to 24 months. We argue that autoencoder performance illustrates concurrent density advantages and disadvantages, in contrast to prior behavioural and computational literature treating such effects independently. Low network error rates signal a configuration advantage for high-density words, while high network error rates signal a triggering advantage for low-density words. This interpretation is consistent with the application of autoencoders in academic research and industry, for simultaneous feature extraction (i.e. configuration) and anomaly detection (i.e. triggering). Autoencoder simulation therefore illustrates how apparently contradictory density and distinctiveness effects can emerge from a common learning mechanism.

1. Introduction

Words with high phonological neighborhood density (i.e. words that sound similar to many other words in the language to which children are exposed) are learned developmentally earlier and remembered and produced more accurately than words with low phonological neighborhood density (Fourtassi, Bian, & Frank, 2018; Hollich, Jusczyk, & Luce, 2002; Stokes, 2014; Storkel, 2004). One way to understand this effect is in terms of long-term auditory priming (e.g. Church & Fisher, 1998). In this account, phonological representations of words heard in child-directed and overheard speech are formed in the child's long-term memory (Port, 2007). These representations may be perceptual, meaning that they are stored without semantic details, or they may be conceptual, meaning that they are stored with semantic details. High neighborhood density words are memorized more easily than low neighborhood density words because high-density words contain sound features that are well represented in existing perceptual and conceptual word memories. The novel high-density word *coal*, for instance, may be acquired through analogy to existing memories including *coat*, *pole*, *cone*, *hole, code*, and *mole* (Church & Fisher, 1998).

One challenge for research in early word learning has been to reconcile evidence of a high-density word learning advantage with contrasting evidence of a high-density word learning *dis*advantage in specific contexts (e.g. Stager & Werker, 1997; Swingley & Aslin, 2007). Swingley and Aslin (2007), for instance, found that children aged 1;6 (one year, six months) struggled to associate phonologically similar labels (e.g. *tog*, neighboring the known word *dog*) to novel objects and reported a learning advantage for distinctive stimuli with no or very few phonological neighbors (e.g. *meb*). One interpretation of this finding is that children may misidentify a novel high-density word as an instance of a known neighbour, particularly in the absence of additional cues to support word leaning, such as a sentence frame or speaker gaze. This behavior is generally adaptive because stored word sound memories and related perceptual mechanisms must be flexible enough to support cross-contextual comprehension on the fly, for instance when a learner encounters a known word in an unfamiliar dialect (Church & Fisher, 1998). Furthermore, the number of minimally different words that young children know and hear regularly in the speech directed to them is limited (Guevara-Rukoz et al., 2018), and this makes it reasonable to classify a novel sound sequence that is very similar to a known word as an instance of that known word instead of as an instance of an unknown word (Swingley & Aslin, 2007).

Overall, then, the evidence suggests that phonological density and phonological distinctiveness support different aspects of word learning. Phonological distinctiveness supports the *triggering* of word learning, in which potential targets of acquisition are identified. Phonological density meanwhile supports lexical *configuration*, or the formation and ongoing fine-tuning of sound memories for these words. These effects have commonly been treated separately, as in the aforementioned studies by Storkel (2004) and Swingley and Aslin (2007), and in related work by Hoover, Storkel, and Hogan (2010) and McKean, Letts, and Howard (2014). Furthermore, there has been a tendency to frame evidence of either a high-density or high-distinctiveness learning advantage as evidence against the opposing effect (e.g. as in Vitevitch & Storkel's, 2013, p. 520, reference to Swingley & Aslin, 2007). The purpose of the current study is to provide a unified framework for understanding apparently contradictory density and distinctiveness effects in early word learning. We use a simple autoencoder neural network to illustrate how these effects can emerge from a common underlying mechanism.

The current study was motivated by Vitevitch and Storkel (2013), who examined neighborhood effects in early word learning by training and testing an autoencoder on a small number of monosyllabic non-words (*N*=60), which were dichotomized into high-density and low-density groups. One novel contribution of the current study is to determine how the high-density advantage reported by Vitevitch and Storkel (2013) scales when using sizeable naturalistic data. In order to make the training data representative of young children's input, we trained an autoencoder on 587,954 word tokens (5497 word types) spoken by 279 caregivers to English-speaking children aged 18 to 24 months. This age range was selected to reflect participants in the aforementioned literature on density and distinctiveness effects (e.g. Storkel, 2004; Swingley & Aslin, 2007). The training data included mono- and multi-syllabic words from all grammatical classes, for instance nouns, verbs, adjectives, and prepositions. To test the trained network, we simulated a MacArthur-Bates communicative development inventory administration (Fenson et al., 2007). Then, to validate network performance, we compared the results of this simulated administration to those from 2292 real administrations involving children aged 18 to 24 months. Note that this validation phase was not possible in prior work using non-words (Vitevitch & Storkel, 2013). In addition to testing the network's ability to represent and output trained words, we also tested the network's ability to generalize and process new, previously untrained words. In all phases, neighbourhood density was modeled continuously, avoiding dichotomization that can reduce statistical power and limit the quality of inferences drawn.

Our interpretation of network performance is informed by our understanding of the application of autoencoders in academic research and industry. Autoencoders are a class of neural networks in which – in three-layer instantiations – input is received in the first layer, compressed in a second 'hidden' layer, and then reconstructed in a third output layer. Autoencoders learn through back propagation, updating between-layer connection weights in order to reduce input-output error.

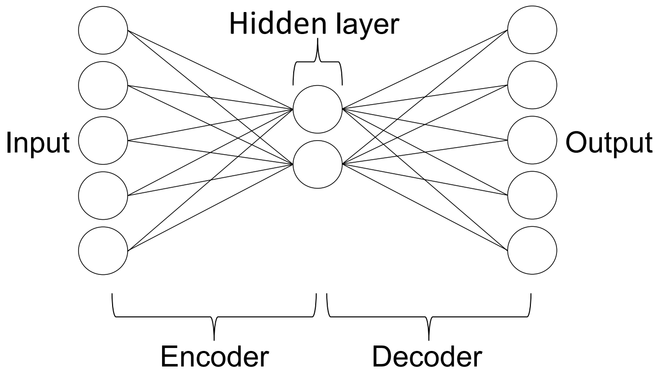

Figure 1. A simplified autoencoder architecture.

Autoencoders show large error when there is a big difference between the input data representation and the output data representation. Importantly, whether or not high network error is undesirable depends on the task at hand. Low error indicates that a given data point has features consistent with the well-represented properties of the previous network input, such as the dominant features in a set of images or the semantic or phonological features common across a set of words. In the context of neighborhood density effects, the low error rate reported by Vitevitch and Storkel (2013) represents a configuration advantage for high-density words. However, high error may be considered advantageous when the purpose of the autoencoder is to detect anomalies. For example, in credit card fraud detection, an autoencoder may be trained on non-fraudulent transactions only, with both non-fraudulent and fraudulent transactions subsequently presented and the latter prompting an increase in error rate. Similarly, in a categorization task simulation, the network may habituate to a set of similar stimuli and de-habituate on presentation of an anomalous stimulus. In each case, high error rates indicate that a novel data point (i.e. a transaction or stimulus) is unlikely to be a member of any trained class. In the context of simulating neighborhood density effects in early word learning, a spike in error rate indicates that a novel string is unlikely to be an instance of any previously trained word. And in this sense, high autoencoder error provides a strong analogy to the triggering advantage for distinctive stimuli observed in human participants (e.g. Swingley & Aslin, 2007).

A broad similarity may be seen between the computational approach used in this study and behavioural paradigms such as the naming task, in which participants must accurately read a word or verbally label a stimulus, or the non-word repetition task, in which participants must accurately repeat a nonsense auditory word stimulus. In each case, lower error rates are taken as evidence of better-memorized properties of the input. However, we want to emphasize that the focus of this report is a simple model of word sound memory configuration and associated triggering effects, rather than an explicit model of word comprehension or production. In addition, we remain agnostic regarding the nature of actual word sound representations, for instance prototypes, exemplars, or hybrids (see Ambridge, 2018, for discussion).

2. Method

2.1. Network specification

A full network specification can be retrieved via the R code hosted on the Open Science Framework repository associated with this project (https://osf.io/2qk5j/). We used the h2o machine learning platform (H2O.ai, 2016) to build an autoencoder with rectified linear unit activation functions, a learning rate of .1, one thousand training epochs, and randomized initial weights. These parameters make our network broadly comparable to that of Vitevitch and Storkel (2013). Our autoencoder had 114 input nodes and 114 output nodes; a number determined through the numerical encoding of words from the training corpus (see section 2.2., *Training*). In a basic sensitivity analysis, we compared networks with 10, 20, and 30 hidden-layer nodes, i.e. with smaller or larger processing resources. Having observed equivalent main effects we settled on a hidden-layer size of 20 nodes.

2.2. Training

The autoencoder was trained on 587,954 tokens (5497 mono- and multi-syllabic unique word types, including all grammatical classes) from child-directed speech from 279 caregivers, directed at American English-speaking children aged 18 to 24 months. These data were retrieved from the Child Language Data Exchange System (CHILDES) using the childesr package in R (MacWhinney, 2000; Sanchez et al., 2018). For each word type we extracted a machine-readable phonological encoding (i.e. a string of 0s and 1s; an example follows) from the pre-embedded Medical Research Council (MRC) dictionary hosted as part of the PyPatPho package (Coltheart, 1981; Grimm & Tulkens, 2015; see https://github.com/RobGrimm/PyPatPho). Only words listed in this database were included in the training inventory. These numerical encodings were generated using PatPho via PyPatPho in Python (Grimm & Tulkens, 2015; Li & MacWhinney, 2002). PatPho converts words into 114-unit binary value vectors on the basis of a range of articulatory features (e.g., voiced, voiceless, front, back, labial, high, lateral, etc.) adopting a syllabic template scheme that accommodates input of varying length and therefore enabling us to model mono- and multi-syllabic words within a parallel architecture. Truncated example PatPho encodings for the words *cat* and *hat* are shown below. Note that encodings were fronted, meaning that word-initial features start at the beginning of the 114-digit vector.

/kæt/ = [0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0...0_{144}]

/hæt/ = [0 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0...0_{144}]

Shading identifies the portion of the vector containing the differences in 0s and 1s that map to the difference in the first phonemes of *cat* and *hat* (i.e. /k/ versus /h/). The subsequent string identity – continuous up to 114 digits – reflects the shared phonemes /æt/ and placeholders supporting the encoding of longer, multi-syllabic words. During training, the encoded child-directed speech corpus was passed to the network defined in section 2.1, *Network specification*.

2.3. Test

After training, we tested the network on a 586-item subset of the trained data that appear in the MacArthur-Bates communicative development inventory, words and sentences version (MCDI-WS; Fenson et al., 2007). The MCDI-WS contains a list of words and phrases and accompanying checkboxes under the response option 'produces'[1]. During real-world administration, caregivers are asked to tick the boxes next to the words that their child is able to say. We accessed the MCDI-WS data using the wordbankr package in R (Braginsky, Yurovsky, Frank, & Kellier, 2018; Frank, Braginsky, Yurovsky, & Marchman, 2017). The test word list was encoded using the process described in section 2.2., *Training*.

For each test word we calculated three independent variables: Phonological neighborhood density, frequency, and length. Following Luce and Pisoni (1998), developmental researchers commonly define phonological neighborhood density as the number of words in a given corpus that can be formed by the addition, substitution, or elimination of a single phoneme in a target word, e.g. *cat* neighbours *hat*, *cot, can,* and *catch*. A limitation of this approach, however, is that many of the words to which young children are exposed are 'lexical hermits' with zero plus/minus one-phoneme neighborhood density. Accordingly, we used a continuous metric of similarity called phonological Levenshtein distance, or PLD20, defined as the mean number of additions, substitutions, or eliminations of phonemes required to change a particular word into its nearest twenty phonological neighbours (Suárez, Tan, Yap, & Goh, 2011, p. 606). PLD20 values for each test word were calculated using all words in the training corpus. A smaller PLD20 indicates greater phonological similarity (i.e. high density).

Frequency and length variables were also included in our statistical model because close association with neighborhood density (i.e. high-density words are typically high

[1] Note that we only tested MCDI-WS words and that MCDI-WS phrases were excluded from our analysis.

frequency and short) makes it important to control statistically for these effects. Previous studies have also reported interactions between these variables. For instance, Storkel (2004) found a significant association between high phonological neighborhood density and early age-of-acquisition for low- but not high-frequency words. In the current study, we used log token frequencies for each test word in the training inventory, and length was measured in number of phonemes. Alternative measures of word length, including number of letters, syllables, or morphemes, are highly correlated and may therefore provide similar results (Lewis & Frank, 2016). We selected the phoneme-based measure given the central interest in this unit of representation in the current study (i.e. as the basis of the PLD20 calculation).

The statistical analysis of test phase error rates was conducted in R (R Core Team, 2016) using the brms package (Bayesian regression models using Stan) (Bürkner, 2017). For all models, likelihood functions were selected on the basis of response variable distribution. In the test phase analysis, we fitted a multiple regression model with a lognormal likelihood, in which autoencoder mean squared error was predicted by word frequency, word length, phonological distance (PLD20), and interactions between PLD20 and word frequency and length (i.e. PLD20*frequency, PLD20*length). We used brms default priors, with each predictor centered and scaled prior to model fitting. This model fitted successfully, with a good number of effective samples, stationery and well-mixing chains, rhats uniformly at 1, and credible posterior predictive checks (see R code for full diagnostics, and the brms package documentation for further description of diagnostic terminology; Bürkner, 2017).

2.4. Validation

Using real words during training and test made it straightforward to compare network performance to data from children. We used the network's test-phase error rates to predict rates of word production among 2292 American English-speaking children aged 18 to 24 months, i.e. matched in age to the training inventory. That is, we compared the results of our simulated MacArthur-Bates communicative development inventory administration to a large database of completed, real-world administrations. This data was retrieved from the wordbank database using the wordbankr package in R (Braginsky et al., 2018; Yurovsky, Frank, et al., 2018; Frank et al., 2017; R Core Team, 2016). We calculated the proportion of children that were able to produce each test word and used this as the dependent variable in a Bayesian regression model in which the by-word mean squared error rates from our autoencoder was the independent variable. We used a gamma family likelihood and brms default priors, and the predictor was centered and scaled for model fitting (see R code for diagnostics).

2.5. Generalization

In this phase, we exposed the trained network to 500 words it had not been trained on and measured the error rates for these items. Generalization-phase words were randomly sampled from the Massive Auditory Lexical Decision (MALD) database (Tucker et al., 2018), and the degree of phonological similarity between each generalization word and words in the training inventory was calculated using the PLD20 metric. The question addressed in this analysis was whether error rates were higher or lower for generalization words that sounded relatively similar or dissimilar to words that the autoencoder had been trained on. We addressed this question using a Bayesian regression model in which generalization word mean squared error rate was predicted by PLD20 and word length in phonemes. We used a skew-normal family likelihood and brms default priors, with predictors again centered and scaled for model fitting (see R code for diagnostics).

3. Results

We begin with the results from the test phase, in which we simulated a MacArthur-Bates communicative development inventory (MCDI-WS) administration on an autoencoder trained on a large corpus of authentic child-directed speech (see Appendix for model summaries). We found main effects for each predictor, which are visualized as posterior probability distributions in Fig. 2. High reconstruction error rates were associated with: (i) Long word length in phonemes (β =0.04; error=0.02; lower 95% credible interval=-0.00; upper 95% credible interval=0.08); (ii) low child-directed speech frequency (β =-0.02; error=0.01; lower 95% credible interval=-0.04; upper 95% credible interval=0.00); and (iii) high phonological Levenshtein distance (PLD20), i.e. low phonological neighborhood density (β =0.18; error=0.02; lower 95% credible interval=0.13; upper 95% credible interval=0.22).

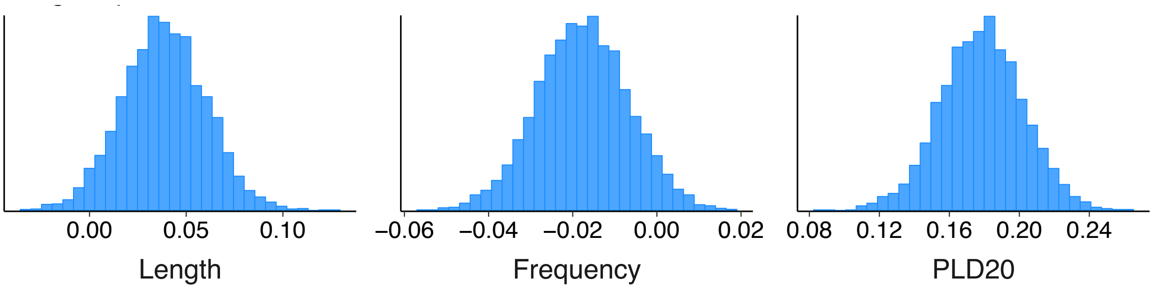

Figure 2. Posterior probability distributions for the beta (β) coefficients representing the association between autoencoder mean squared error and; (i) word length (in phonemes), (ii) log child-directed speech frequency, and (iii) phonological Levenshtein distance (PLD20).

We also found evidence of a higher-order interaction between PLD20 and word frequency (β =-0.04; error=0.01; lower 95% credible interval=-0.07; upper 95% credible interval=-0.02). This indicates that the association between high neighborhood density and low error rate was particularly strong for low frequency words, with high frequency nullifying the PLD20 effect. No higher-order interaction was observed between word length and PLD20 (β =-0.01; error=0.01; lower 95% credible interval=-0.02; upper 95% credible interval=0.01).

During the subsequent validation phase, we used the error rates from our simulated MCDI-WS administration to predict proportions of MCDI-WS word production among 2292 American English-speaking children matched in age to the training inventory (i.e. 18-24 months). We found a negative trend, with words with higher autoencoder error rates produced by a smaller proportion of children (*β* =-0.03; error=0.03; lower 95% credible interval=-0.09; upper 95% credible interval=0.02).

Finally, during the generalization phase, we exposed the trained autoencoder to a randomly sampled inventory of 500 previously unseen words that varied in phonological similarity to words in the training inventory. Higher error rates were observed for high-PLD20 (i.e. low-density) words when controlling for the effect of word length (*β* =0.02; error=0.00; lower 95% credible interval=0.01; upper 95% credible interval=0.02).

4. Discussion

This study used an autoencoder neural network to simulate phonological neighborhood density and distinctiveness effects observed in early word learning. One contribution of this study was to determine how the results of Vitevitch and Storkel (2013) scaled when using sizeable naturalistic training and test data, avoiding data dichotomization, and incorporating validation against real world data. We trained a three-layer autoencoder using a large corpus of child-directed speech before simulating a communicative development inventory administration at test and then comparing network performance to that of children who were age-matched to the training data (i.e. 18-24 months). Lower reconstruction error rates were observed for words that sounded similar to many other words in the child-directed speech on which the autoencoder was trained. This effect was separable from the effects of word frequency and word length, which also tended in the expected directions given the existing behavioral data. That is, lower error rates were observed for high frequency words and for short words (Braginsky, Yurovsky, Marchman, & Frank, 2018). Despite the extreme simplicity of our network, we were therefore able to simulate the high phonological neighborhood density configuration advantage reported behaviorally (e.g. Fourtassi et al., 2018; Hollich, Jusczyk, & Luce, 2002; Stokes, 2014; Storkel, 2004). We also reported a higher-order interaction between word frequency and phonological distance. As demonstrated behaviorally by Hollich et al. (2002) and Storkel (2004), we found that high frequency nullified the high phonological neighborhood density advantage, with amplified error rates for low-frequency, low-density words.

In network validation, we used test-phase error to predict word production rates among 2292 children. Despite a credible interval including zero – indicating that zero may be the true value of the effect – we observed a negative trend in which fewer children produced words that the autoencoder had difficulty representing and reconstructing at test (β =-0.03). Finally, we examined the network's ability to generalize to previously unseen data and found an advantage for words with low PLD20 (i.e. high density) relative to the training corpus. That is, the autoencoder was better able to represent and reconstruct novel words that sounded similar to trained words than novel, phonologically anomalous words. Broadly similar results have been reported behaviorally by Schwartz and colleagues, who found that children were more likely to learn to successfully produce a novel word if that word contained IN-sounds – i.e. sounds that the child had previously produced – than if it contained previously unattested OUT-sounds (Schwartz & Leonard, 1982; Schwartz, Leonard, Frome Loeb, Swanson, & Loeb, 1987; see also Storkel, 2006).

High neighborhood density is associated with low network error because the encodings of phonologically similar words exhibit similar patterns (i.e. comparable series of 0s and 1s; see the *cat* and *hat* example in section 2.2., *Training*) that are better represented across the network during dimensionality reduction, a process sometimes termed a *conspiracy effect* in machine learning (Rumelhart, McClelland, and the PDP Research Group, 1986). This makes it possible to reconstruct high phonological neighborhood density words more accurately, as reflected in low error rates during training, test and generalization. For instance, exposure to the words *coat*, *pole*, *cone*, *hole, code*, and *mole* prompts changes in the connection weights that support the reconstruction of the novel neighbor *coal*. As the autoencoder is forced through the hidden layer bottleneck (see Fig. 1) to extract dominant input properties, generalization to a novel word exhibiting features orthogonal to those previously experienced is inhibited, as reflected by high reconstruction error rates for phonologically distinctive, high PLD20 words.

In our view, a real world parallel to the computational mechanism described above is the cognitive process of long-term auditory priming (e.g. Church & Fisher, 1998). In this account, representations of direct and indirect spoken word exposures are stored in long-term memory (Port, 2007). These representations are initially perceptual rather than conceptual in nature and may be formed implicitly in the absence of semantic information, much like the representations formed by our network. Children are sensitive to the degree of similarity between stored perceptual representations and are able to use this sensitivity to identify (e.g. in the head-turn preference procedure) word sounds that occur at high-probability in their native language (Fourtassi et al., 2018; Jusczyk, Luce, & Charles-Luce, 1994). Novel high-density target words comprising phonological features consistent with existing perceptual memory traces may be held in memory more easily during initial processing (Gathercole, Frankish, Pickering, & Peaker, 1999; Hoover et al., 2010), and this supports the formation of long-term, perceptual and conceptual memory traces that are well detailed and robust to forgetting (Metsala & Walley, 1998; Sosa & Stoel-Gammon, 2012; Storkel, 2004; Walley, Metsala, & Garlock, 2003). Learners may increasingly use their awareness of high-probability word sounds, as well as their related aptitude in producing such sounds, to generalize readily to novel though phonologically familiar words, as in the aforementioned IN-sound/OUT-sound studies of Schwartz and colleagues (Schwartz & Leonard, 1982; Schwartz et al., 1987; see also Storkel, 2006). Low-density words are in general difficult for young children to acquire because there exist few similar stored word representations – whether perceptual or conceptual – from which to generalize.

In the introduction we noted a tendency in the prior literature to treat density and distinctiveness effects separately, and to frame evidence of either a high-density or high-distinctiveness learning advantage as evidence against the opposing effect (e.g. Storkel, 2004; Swingley & Aslin, 2007; Hoover et al., 2010; McKean et al., 2014; Vitevitch & Storkel, 2013). In contrast to this approach, the second contribution of this study is to provide a unified framework for understanding density and distinctiveness effects in early word learning. To do this, we want to emphasize that autoencoder neural networks perform both feature extraction and anomaly detection in parallel. In this sense would be inaccurate to suggest that high autoencoder error rates for low-density words provide an analogy to learning deficits in children (Vitevitch & Storkel, 2013). Whereas low network error rates may indeed be understood as exposure to high-density words prompting a conspiracy effect supporting lexical configuration, high autoencoder error signals the detection of an anomalous target word comprising phonological features inconsistent with those previously trained. This latter effect – i.e. computational anomaly detection – parallels the triggering advantage observed for low-density words in children (e.g. Stager & Werker, 1997; Swingley & Aslin, 2007), which itself may be decomposed into attention- or curiosity-based learning advantages (Twomey & Westermann, 2017; we note that additional learning mechanisms conceivably dependent on the fundamental triggering mechanism simulated form no part of our model). Autoencoder neural networks therefore provide a neat computational analogy to both the density advantages and the distinctiveness advantages observed in behavioral studies of early word learning. Triggering effects may be seen as the advantageous by-product of long-term auditory priming (or a conspiracy effect), which itself supports lexical configuration. These effects can be simulated in parallel within a single autoencoder employing common algorithms and parameter values. In this way, autoencoder simulation illustrates how apparently contradictory density and distinctiveness advantages emerge from a common cognitive mechanism.

The current study demonstrates neighbourhood density and distinctiveness effects similar to those observed in young children in the absence of semantic and pragmatic information. This illustrates the crucial role that raw auditory word similarity plays in the formation of the early lexicon. It is important to emphasize, however, that high phonological neighborhood density is just one of many factors supporting early word learning, including high exposure frequency, high concreteness, high relevance to babies and infants, and alternative sound variables including phonotactic probability, i.e. the probability of phoneme co-occurrence (Braginsky, Yurovsky, Marchman, et al., 2018; Jones & Brandt, 2018; see section 4.1, *Limitations*, for discussion of phonotactic probability). The current study, for instance, accorded with prior behavioral work in reporting that the high neighborhood density effect was nullified by high exposure frequency (e.g. Hollich et al., 2002; Storkel, 2004); a finding that suggests an apparent primacy of word-level frequency effects relative to word sound characteristics. It is therefore expected that if a child hears a target word frequently enough, or if that target word is, for instance, highly concrete or highly relevant to the child, then the implicit generalization preference for words with familiar phonological properties may be nullified.

4.1 Limitations

Computational cognitive modeling requires researchers to make numerous decisions, from the overall model type used (e.g. a neural network or Bayesian network) to fine-grained details regarding parameters (e.g. priors, network learning algorithm and rate, number of training epochs, etc.). Inevitably, then, some readers may question particular choices we made. One particular point of concern may be our decision to use an autoencoder rather than a recurrent neural network or long short-term memory network, given that recurrent architectures are so commonly used in natural language processing research. The rationale for our choice of architecture was twofold. First, an autoencoder was used in the work by Vitevitch and Storkel (2013) that inspired this study, and replication with naturalistic data necessitated the use of the same architecture. Second, autoencoders are a somewhat distinctive branch of architecture in the sense of performing parallel feature extraction and anomaly detection. This choice of architecture was therefore essential to our aim of illustrating how apparently contradictory behavioural evidence of both density and distinctiveness advantages can be explained in terms of a common mechanism. We have made all of our data and code fully available online, and researchers are welcome to access this material to test alternative configurations of network or encoding approaches.

Another potential limitation of this report is the exclusion of alternative predictor variables, perhaps most importantly phonotactic probability. High positive correlation between neighborhood density and phonotactic probability may cause multicolinearity (Storkel, 2004; Storkel & Lee, 2011), which distorts results by changing the magnitude or the direction of estimates, or by inflating estimate errors. While it is possible to tease apart the effects of neighborhood density and phonotactic probability in controlled experimental settings (e.g. Storkel & Lee, 2011), this is usually not possible when working with naturalistic data or communicative development inventory data (see Storkel, 2004, with respect to MacArthur-Bates data). In this case, the safest way to address multicolinearity risk is to exclude the variable of least interest from the regression model. For us, given our central interest in neighborhood density effects, this meant omitting phonotactic probability. However, as one anonymous reviewer commented, this makes it impossible to determine the contribution of phonotactic probability to the results presented. We would like to re-emphasize that all our code and data can be accessed via the project repository accompanying this paper, and that researchers with a primary interest in sub-lexical phonotactic probability effects rather than the word-level neighborhood density effects covered in this study are welcome to modify these materials.

5. Conclusion

High phonological neighborhood density has been associated with both advantages (Storkel, 2004) and disadvantages (Swingley & Aslin, 2007) in behavioral studies of early word learning. We explored these effects using an autoencoder neural network in conjunction with corpus and communicative development inventory data. We suggested that the widely reported high-density advantage is explicable in terms of exposure to a phonological neighborhood prompting a natural conspiracy effect; a process termed long-term auditory priming in the behavioural literature (e.g. Church & Fisher, 1998). We then noted that high phonological distinctiveness supports word learning by reducing the risk of mis-processing novel words as known words in competitive learning environments. Autoencoder modeling encourages us to think of these apparently contradictory effects as emerging from a common mechanism.

References

Ambridge, B. (2018). *Against stored abstractions: A radical exemplar model of language acquisition (Unpublished pre-print)*. Retrieved from https://ling.auf.net/lingbuzz/004131

Braginsky, M., Yurovsky, D., Frank, M., & Kellier, D. (2018). wordbankr: Tools for connecting to wordbank, an open repository for developmental vocabulary data. Retrieved from https://github.com/langcog/wordbankr

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2018). Consistency and variability in word learning across languages. https://doi.org/10.31234/osf.io/cg6ah

Bürkner, P.-C. (2017). brms: Bayesian Regression Models using "Stan." CRAN repository. Retrieved from https://cran.r-project.org/web/packages/brms/index.html

Church, B. A., & Fisher, C. (1998). Long-term auditory word priming in preschoolers: Implicit memory support for language acquisition. *Journal of Memory and Language*. https://doi.org/10.1006/jmla.1998.2601

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. https://doi.org/10.1080/14640748108400805

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates communicative development inventories: User's guide and technical manual* (2nd ed.). Baltimore, MD: Brookes.

Fourtassi, A., Bian, Y., & Frank, M. C. (2018). Word learning as network growth: A cross-linguistic analysis. Retrieved from http://langcog.stanford.edu/papers_new/fourtassi-2018-cogsci.pdf

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(03), 677–694. https://doi.org/10.1017/S0305000916000209

Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/0278-7393.25.1.84

Grimm, R., & Tulkens, S. (2015). PyPatPho: A phonological pattern generator. GitHub repository. Retrieved from https://github.com/RobGrimm/PyPatPho

Guevara-Rukoz, A., Cristia, A., Ludusan, B., Thiollière, R., Martin, A., Mazuka, R., & Dupoux, E. (2018). Are words easier to learn From infant- than adult-directed speech? A quantitative corpus-based investigation. *Cognitive Science*. https://doi.org/10.1111/cogs.12616

H2O.ai. (2016). R Interface for H2O, R package. GitHub repository.

Hollich, G., Jusczyk, P. W., & Luce, P. A. (2002). Lexical neighborhood effects in 17-month-old word learning. *Proceedings of the 26th Annual Boston University Conference on Language Development*, (January), 314–323.

Hoover, J. R., Storkel, H. L., & Hogan, T. P. (2010). A cross-sectional comparison of the effects of phonotactic probability and neighborhood density on word learning by preschool children. *Journal of Memory and Language*, *63*(1), 100–116. https://doi.org/10.1016/j.jml.2010.02.003

Jones, S. D. &, Brandt, S. (2018). *Do children really acquire dense neighborhoods?* Manuscript submitted for publication.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*. https://doi.org/10.1006/jmla.1994.1030

Li, P., & MacWhinney, B. (2002). PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments, & Computers*, *34*(3), 408–415. https://doi.org/10.3758/BF03195469

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

McKean, C., Letts, C., & Howard, D. (2014). Triggering word learning in children with Language Impairment: The effect of phonotactic probability and neighbourhood density. *Journal of Child Language*, *41*(6), 1224–1248. https://doi.org/10.1017/S0305000913000445

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk, Vol 1: Transcription format and programs. The CHILDES project: Tools for analyzing talk, Vol 1: Transcription format and programs (3rd ed.).* (3rd ed.). New York: Psychology Press (Taylor and Francis group).

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 89–120). Mahwah, NJ.

Port, R. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology*. https://doi.org/10.1016/j.newideapsych.2007.02.001

Python Software Foundation. (2013). Python language reference. *Python Software Foundation*. https://doi.org/https://www.python.org/

R Core Team. (2016). R. *R Core Team*. https://doi.org/3-900051-14-3

Rumelhart, D. E., McClelland, J. L. & the PDP Research Group (1986). Parallel distributed Processing: Explorations in the Microstructure of Cognition (Volume 1: Foundations). Cambridge, MA: MIT Press.

Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K., Yurovsky, D., & Frank, M. (2018). childes-db: a flexible and reproducible interface to the Child Language Data Exchange

System. Retrieved from https://psyarxiv.com/93mwx

Schwartz, R. G., & Leonard, L. B. (1982). Do children pick and choose? An examination of phonological selection and avoidance in early lexical acquisition. *Journal of Child Language*. https://doi.org/10.1017/S0305000900004748

Schwartz, R. G., Leonard, L. B., Frome Loeb, D., Swanson, L. A., & Loeb, D. M. (1987). Attempted sounds are sometimes not: an expanded view of phonological selection and avoidance. *Journal of Child Language*. https://doi.org/10.1017/S0305000900010205

Sosa, A. V., & Stoel-Gammon, C. (2012). Lexical and phonological effects in early word production. *Journal of Speech Language and Hearing Research*, *55*(2), 596. https://doi.org/10.1044/1092-4388(2011/10-0113)

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381–382. https://doi.org/10.1038/41102

Stokes, S. F. (2014). The impact of phonological neighborhood density on typical and atypical emerging lexicons. *Journal of Child Language*, *41*(3), 634–657. https://doi.org/10.1017/S030500091300010X

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*(2), 201–221. https://doi.org/10.1017/S0142716404001109

Storkel, H. L. (2006). Do children still pick and choose? The relationship between phonological knowledge and lexical acquisition beyond 50 words. *Clinical Linguistics and Phonetics*, *20*(7–8), 523–529. https://doi.org/10.1080/02699200500266349

Storkel, H. L., & Lee, S. (2011). The independent effects of phonotactic probability and neighborhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, *26*(2), 191–211. https://doi.org/10.1080/01690961003787609

Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-011-0078-9

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, *54*(2), 99–132. https://doi.org/10.1016/j.cogpsych.2006.05.001

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2018). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1056-1

Twomey, K. E., & Westermann, G. (2017). Curiosity-based learning in infants: A neurocomputational approach. *Developmental Science*, (September), 1–13. https://doi.org/10.1111/desc.12629

Vitevitch, M. S., & Storkel, H. L. (2013). Examining the acquisition of phonological word forms with computational experiments. *Language and Speech*, *56*(4), 493–527. https://doi.org/10.1177/0023830912460513

Walley, A. C., Metsala, J. L., & Garlock, V. M. (2003). Spoken vocabulary growth: Its role in the development of phoneme awareness and early reading ability. *Reading and Writing: An Interdisciplinary Journal*, *16*(1), 5–20. https://doi.org/10.1023/A:1021789804977

Yermolayeva, Y., & Rakison, D. H. (2014). Connectionist modeling of developmental changes in infancy: Approaches, challenges, and contributions. *Psychological Bulletin*. https://doi.org/10.1037/a0032150

Appendix

Model summaries.

Table A1

Test phase model summary showing term (main effects, interactions, and family specific parameters), estimate, standard error (Std. error), and lower (L) and upper (U) 95% confidence intervals (CI). Model formula: *Mean reconstruction error ~ Length + Frequency + PLD20 + PLD20 * Length + PLD20 * Frequency*.

Term (main effects)	Estimate	Std. error	L-95% CI	U-95% CI
Intercept	-3.38	0.01	-3.4	-3.36
PLD20	0.18	0.02	0.14	0.22
Length	0.04	0.02	0	0.07
Frequency	-0.02	0.01	-0.04	0
Term (interactions)	Estimate	Std. error	L-95% CI	U-95% CI
PLD20: Length	-0.01	0.01	-0.02	0
PLD20: Frequency	-0.04	0.01	-0.06	-0.02
Term (family specific parameters)	Estimate	Std. error	L-95% CI	U-95% CI
Sigma	0.23	0.01	0.22	0.24

Table A2

Validation phase model summary showing term (main effects and family specific parameters), estimate, standard error (Std. error), and lower (L) and upper (U) 95% confidence intervals (CI). Model formula: *Produces (%) ~ Mean squared error*.

Term (main effects)	Estimate	Std. error	L-95% CI	U-95% CI
Intercept	-1.21	0.03	-1.25	-1.16
Mean squared error	-0.03	0.03	-0.08	0.02
Term (family specific parameters)	Estimate	Std. error	L-95% CI	U-95% CI
Shape	1.98	0.11	1.8	2.16

Appendix continued.

Model summaries.

Table A3

Generalization phase model summary showing term (main effects and family specific parameters), estimate, standard error (Std. error), and lower (L) and upper (U) 95% confidence intervals (CI). Model formula: *Mean reconstruction error ~ PLD20 + Length*.

Term (main effects)	Estimate	Std. error	L-95% CI	U-95% CI
Intercept	-0.01	0	-0.01	0
PLD20	0.02	0	0.01	0.02
Length	0	0	0	0.01
Term (family specific parameters)	Estimate	Std. error	L-95% CI	U-95% CI
Sigma	0.02	0	0.02	0.02
Alpha	1.83	0.4	1.19	2.45