A Multi-Relationship Language Acquisition Model for Predicting Child Vocabulary Growth

Andrew Roxburgh, Floriana Grasso, and Terry R Payne

Department of Computer Science, University of Liverpool, Liverpool, L69 3BX
Corresponding author: a.roxburgh@liverpool.ac.uk

Abstract. If we can predict the words a child is likely to learn next, it may lay the foundations for developing a tool to assist child language acquisition, especially for children experiencing language delay. Previous studies have demonstrated vocabulary predictions using neural network techniques and graph models; however, individually these models do not fully capture the complexities of language learning in infants. In this paper, we describe a multi-relationship-layer predictive model, based on a graph neural network. Our model combines vocabulary development over time with quantified connections between words calculated from fifteen different norms, incorporating an ensemble output stage to combine the predictions from each layer. We present results from each relationship layer and the most effective ensemble arrangement.

Keywords: graph neural networks · language acquisition

1 Introduction

The acquisition of language and communication skills during early years plays a crucial role in the overall cognitive and social growth of children, such that any interruption or delay can have far reaching consequences to language development and educational attainment in later years. Language impairment, where a child's language abilities are insufficient for their next stage of cognitive, educational, and social development, has been demonstrated to impede the child's development from an early age, and without proper support they can fall behind and fail to catch up with their peers [1]. *Developmental Language Disorder* (DLD), which is a condition whereby a child's language development is delayed or disordered for no clear reason, affects 6.44% of all UK children [2], and is the most prevalent childhood disability, requiring specialist support in order for affected children to learn and communicate to the very best of their ability [3]. DLD has been linked with lower academic achievement, lower employment and poor mental health [4]. Even in neurologically 'typical' children, factors such as their communication environment and family circumstances can have an effect on their language development, and research has shown that delayed communication skills can lead to adverse learning outcomes several years later [5].

To establish whether a child's language is developing normally, standardised tools are used such as the MacArthur-Bates Communicative Development Inventory (CDI) [6,7]. The CDI consists of a series of questions and checklists designed

to assess vocabulary comprehension, production, gesture use, and early grammar. It is usually accompanied by a Family Questionnaire that is compiled by the child's primary carer or by a researcher. By comparing the answers against national norms, it is possible to establish whether the child is developing normally or could be delayed.

Originally designed as a paper instrument, online web or mobile platform versions offer novel research prospects such as the capability to offer recommendations and direction to parents regarding the words they should teach their child next.

A simple but naive way of achieving this would be by referencing the *Age-of-Acquisition* norms [8] that identify words that typical children would acquire at a similar age. However, children tend not to learn the same words at the same rate or age, and thus a more tailored approach is required, based on the child's current knowledge. This predictive technique could be used to inform the child's primary carer of 'candidate' words to emphasise when teaching language [9], and form the basis of a novel language intervention tool. In this paper we present a novel approach for predicting a child's language acquisition by utilizing Spatio-Temporal Graph Neural Networks (STGCN), which aims to improve upon the existing literature in terms of accuracy. We evaluate the viability and efficacy of using such a network, utilising published lexical datasets; and illustrate how this approach is worthy of further investigation.

2 Existing work

A seminal work on the prediction of word acquisition by young children based on their current vocabulary, Beckage, Mozer & Colunga [10] explored the use of conditional probabilities by examining the CDI questionnaire data of 77 subjects over a 1-year period at monthly intervals. By using a network growth technique, they built three different models based on calculating the conditional probability of a word being learnt within the next month using different approaches, given words that had been learnt overall and in the previous month. They found that the accuracy of predictions could be enhanced by increasing the temporal resolution of the data (e.g. more frequent than monthly intervals) or by including more meaningful connections between words in the predictive model.

Other work has looked at the use of Artificial Neural Network (ANN) models for predicting the probabilities of word acquisition over a subsequent month [11]. ANNs have a long history of use in early learning research including language modelling, and have proven themselves to be excellent statistical learning tools. A number of different neural-network based predictive models were investigated using various qualitatively different sources of information as inputs [10]. All of these models augmented an initial set of 6 inputs representing demographic information about the child. One model used a representation of the child's current vocabulary, as indicated by the answers provided by their parents to the CDI questionnaire, consisting of an additional 677 inputs. A different model utilised a representation of the semantic features of words in the child's vocabulary,

based on the McRae feature norms [12], through 30 additional inputs. Other models considered the phonological composition of the child's productive vocabulary (represented by 37 additional inputs), or representations that captured the production of words within specific categories of the CDI questionnaire (22 additional inputs). Other studies exploited a Word2Vec [13] based representation of the child's productive vocabulary that combined vectors in a high-dimensional linguistic space and comprising 200 inputs.

Beckage et. al. [11] also explored the use of ensemble models to determine if some language representations were unnecessary or if the combination of multiple representations could improve the model's predictability. From these studies, they observe that: (i) a child's existing vocabulary and demographic information significantly affect their future vocabulary development; (ii) the specific words a child knows are valuable in forecasting their future vocabulary growth; (iii) the model that considered a child's current vocabulary performed better than one relying solely on demographic data; and (iv) the words in a child's vocabulary contain valuable information besides their age and current vocabulary size. They also noted that models based on semantic features and phonology were less effective than those models based on child demographics and current vocabulary, as they don't meaningfully combine the child's existing vocabulary knowledge.

3 Child Vocabulary as a Multi-Relationship Graph

Child vocabulary growth has been modelled in the literature using a variety of network-based methods. Graphs have been used when modelling vocabulary growth over time [14], whereas neural networks were used when attempting to model the way that a brain acquires language [15]. These models exploit the fact that a typical vocabulary consists of a collection of words that are inherently connected with each other, and as such can be easily represented as a network. Typically, words are represented as nodes, with edges representing some inter-word relationship (e.g. Fig. 1 shows a semantic network that focuses on the word *water*). In this model, each node of the graph represents a word and incorporates a feature vector, which contains information about the state or features of the word. Each node is also associated with a state representation of the child's level of knowledge regarding that node: 1) a child may understand a word without production; 2) a child may produce a word without meaning, 3) a child may both understand and produce a word, or 4) a child may have no knowledge of a word. While cognitive nuances of word knowledge extend beyond these four discrete states, for the purpose of analyzing a child's vocabulary, they serve as easily observable and universally understood indicators.

In our study, we enhanced the model by integrating multiple relationships that are effectively superimposed upon each other as in layers. In our structure, the nodes, representing words, are shared across all layers, while each layer exhibits distinct edge configurations (see Fig.2 for a visualisation). To incorporate a new observation into the model, the nodes' feature vectors are suitably adjusted to capture word knowledge at the respective time period.

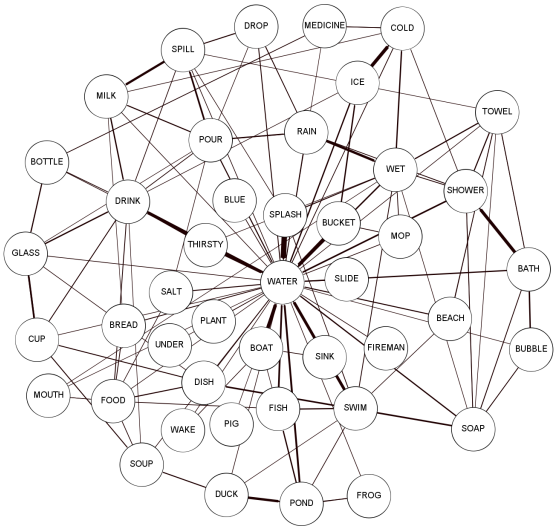

Fig. 1. Simplified example of a vocabulary graph: word-association graph focused on *water*. Edge and node features not shown.

As to which relationships to include in the model, we excluded those focusing on words from an adult viewpoint, and we concentrated on those more aligned with an infant's cognitive perspective, and grounded in the literature. Relationships commonly used in language research include *Semantic Feature norms* (ratings of the attributes or characteristics of words that provide information about their meaning in context), *Word Association data* (for a given cue word, the target word that a person immediately thinks of next), *Phonological Similarity data* (the degree to which words sound similar when spoken), and Psycholinguistic norms such as: *Imageability* (the ease with which a word can be mentally visualised); *Concreteness* (the tangibility of a word as opposed to an abstract concept, e.g. 'chair' is more concrete than 'time'); *Familiarity* (rating a word based on how commonly it is used in everyday speech); and *Word Length* (a measure of how difficult a word is to remember or say). From a cognitive research perspective, *Sensorimotor Norms* allow to compare words from the conceptual point of view of children at the earliest development stage, when they learn to use their senses to build an understanding of the world and use motor movements (grasping, sucking, touching) to interact with it.

Specifically, for our model, we chose: Nelson *et al.* association norms [16], which are utilised to construct a layer that accounts for the associative relationships between words in human memory; the semantic feature production norms by McRae *et al.* [12] and those by Buchanan *et al.* [17] to measure the similarity of meaning between two given words. We also incorporate a measure of Phonological Similarity, based on IPA phonemes, which are extracted from the BEEP phonetic dictionary [18]. Finally, we use the Lancaster Sensorimotor Norms [19] which evaluate English words based on six perceptual modalities: touch, hearing,

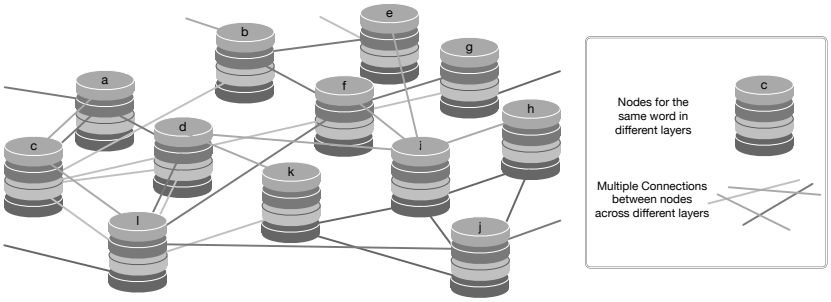

Fig. 2. A Multi-relationship vocabulary graph structure.

smell, taste, vision, and interoception), and five action effectors: mouth/throat, hand/arm, foot/leg, (head excluding mouth/throat), and torso.

Both McRae *et al.* and Buchanan *et al.* data is published along with cosine similarity matrices, allowing for direct representation of connection strengths between words. To similarly adapt the Lancaster norms to our model's structure, a weighted adjacency matrix was created for all possible word pairs within each category. This weight was calculated by normalizing the product of each pair's scores, resulting in a strong connection between words with high scores in the same category and weak connections between dissimilar words. Self-loops are given a weight of 1.0. The Phonological Similarity model was constructed by decomposing each word into its constituent IPA phonemes, derived from the BEEP, and computing, for every pair of words, similarity scores based on the Jaccard similarity metric, allowing us to create an adjacency matrix.

For all models, we used the adjacency matrices to define graph edges, and created a list of nodes by de-duplicating the edges list. The nodes' feature vectors represent the level of knowledge that a child has of the word. This resulted in a collection of graphs $G_n = (V_n, E_n)$ for each category, where V represents the vertices (i.e. nodes) and E represents the edges. A node list was created for every observation in the data and was populated with the corresponding observed data. These node lists were then combined to form a time series. The edge lists were processed by combining each edge list with each node list in the time series, resulting in the creation of a time series of graphs for each of the 15 models. This was used as an input to our model, as explained in the following Section.

4 Model Selection

Our aim is to predict a child's future vocabulary based on the child's past and existing knowledge. Given our series of graphs representing different relationships between words in a vocabulary, we can embed the nodes with feature vectors representing the child's current knowledge of the word (Figure 3). Given that each node has been classified as being 'understood', 'understood and spoken',

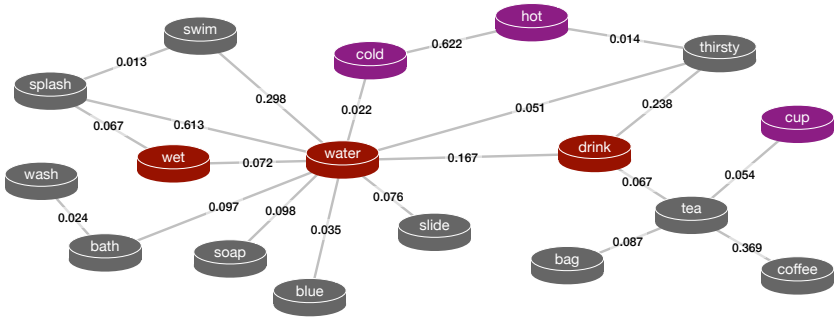

Fig. 3. Example graph portion representing words and semantic connections between them. Edges show strength of connection (in this case semantic relatedness). Node colours represent word's feature vectors: *wet*, *water* and *drink* are known and understood by the child; *cold*, *hot*, and *cup* are either known or understood but not both, and the remaining nodes represent words that are completely unknown to the child.

'spoken but not understood', or 'unknown', a classifier is required that can reclassify the nodes based on their features, connections and history.

Graph Neural Networks (GNN) efficiently apply machine learning to graph-structured data [20]. GNNs process the input graph by taking each node in turn and aggregating information from its neighboring nodes and edges, updating the representations of the nodes in each iteration, until a final representation for each node is obtained. These node embeddings encode the structural and feature data of themselves, their neighbours, and ultimately of all other nodes in the graph, and can then be used in further operations such as edge prediction, classification, labelling, feature prediction and more. In our application we classify the nodes to determine the probability that the words that they represent are 'known'. *Graph Convolutional Networks* [21] are a variety of GNN that attempt to apply a convolution operation to graphs, in a similar manner to traditional Convolutional Neural Networks (CNNs). The type of GCN is determined by the convolution filtering method and is either Spectral (where the convolutions take place in the Fourier domain) or Spatial (in the spatial domain). Following Kipf & Welling [22], we use a technique that bridges the two methods - it uses spectral graph convolutions, but with some simplifications to reduce the processing overhead that comes with computing a Fourier transform of a graph.

An extension of the GCN model is the *Spatio-Temporal Graph Convolutional Network* (STGCN), which considers features of a GCN as a function of both space and time. They have been shown to work well in problems of traffic prediction [23] where the aim is to predict the traffic speeds, given information from sensors on other roads. The data from a road network with traffic sensors can be described as a graph, with the sensors serving as nodes, direct routes between sensors as edges, and the distance between sensors as edge weights. Each node may have features such as vehicle speed or number of passing vehicles. By taking the history of the sensor data into account as well as the relationship of the nodes

to each other, features of particular sensors can be predicted accurately and efficiently [24]. This scenario is analogous to our word prediction problem - nodes representing words rather than sensors, edges representing inter-word relationship strength rather than distance between sensors, and features representing the probability of increased word knowledge rather than vehicle speed. By combining graph convolutional operations with temporal convolutional operations, STGCNs are able to model the dependencies between nodes in a dynamic graph structure over time, making it suitable for forecasting the relationships between nodes at a future point in time, based on recent history as well as current state. This makes it suitable for forecasting the future state of a child's vocabulary, given current and past states of the vocabulary and the relationships between words.

5 Methodology

We have developed a STGCN-based model using Python and *Stellargraph* [25], a software library built on *Tensorflow* [26] which facilitates the construction of graph-based machine learning models. Our full model consists of 15 relationship layers, each of which is a separate STGCN model that has been individually trained and executed. Some nodes in these relationship layers may not have connections as they have no meaningful associations with other words. When a new prediction is required, a vector representing the child's current vocabulary is used to populate the feature vectors of each node on each relationship layer - indicating that certain new words have been learned. The GCN classifier, in conjunction with the STGCN's spatial-temporal block, is then applied to these input graphs to re-classify the 'unknown' nodes, from which we can determine the words that are likely to be influenced the most by its neighbours, and so may be learned next. This produces a list of candidate words from each relationship layer, from which the most likely ones can be determined.

5.1 Assumptions and data preparation

Observational Data. Our observational dataset consists of item-level CDI Survey responses extracted from all available forms in English downloaded from Wordbank [27], combined with additional data collected via volunteers through our website. We have chosen only data for which there are longitudinal sequences, to enable the STGCN algorithm's spatial-temporal block to train on temporal data. Words were converted to our standardised vocabulary to allow for dialect differences. The overall data consisted of 718 observations (i.e. vocabulary inventories), with 150 test subjects, each providing between four and six consecutive observations.

Due to the nature of human-collected data about human behaviour, inevitably there will be errors present. For instance, a child may be observed at one time period as understanding, but not producing, a particular word, e.g. the child may appear to understand 'bath' by going to the bathroom when a

parent says it. On a subsequent observation the child may use the word 'bath', but use it incorrectly to refer to, say, a bath toy kept in the bathroom. Or, the parent may not be aware that a child knows a particular word. The child may have used the word at a grandparents' home for instance. Parents can change their mind if they believe that a child has said and/or understood a word, but then they realise it is not so. Finally there is the added complication of correctly understanding words produced by a toddler, which can often be far from clear. Such data presents challenges, especially in a smaller dataset, where errors can have a bigger negative effect on the model. Given that errors involving words dropping out of the observed vocabulary could adversely affect the ability to accurately train a classifier on the data, we remove contradictory data by generating two datasets: an **optimistic dataset**, in which we assume that children have continued to understand the word in subsequent observation periods, and a **pessimistic dataset** where we assume that they are false observations and that the children did not in fact understand the word during the first observation.

Relationship Datasets. Arguably, standardizing data is one of the biggest challenges when combining multiple independent sources of language data into one model accounting for synonyms ('rabbit' and 'bunny'), multiple dialects ('mommy', 'mummy', 'mom', 'mum', 'ma', 'mama') and international spelling variations ('colour' and 'color'). In these cases we renamed the words to match our own standardised vocabulary. To address the issue of homographs (words with multiple meanings, like 'back' or 'drink'), we maintain standardization by appending a label to ambiguous words (e.g. 'drink' becomes 'drink(beverage)' and 'drink(verb)'). For a child, certain words may hold different meanings compared to adults. As an example, the idea of 'fish' being a food and 'fish' being an animal are typically treated as distinct concepts for children, whereas for adults 'fish' is understood as both a food and an animal at the same time. Again these words are appended with a context-appropriate label. We created a Python script to simplify the labelling and transforming process, necessary for handling all 15 relationship datasets and all observational data.

After finishing the pre-processing of the input data, the data representing the structure of input graphs was generated. For each relationship model, this included Edge data, represented by an adjacency matrix, and a time-series of Node lists, each depicting the state of nodes at a particular observation and featuring a feature vector indicating the child's understanding of the word at that time. This comprehension attribute was assigned a starting value from one of four levels, reflecting the child's knowledge of that word at the given observation. (0.0 representing no comprehension, 0.3 representing production without understanding, 0.6 representing understanding but no production, and 1.0 representing full comprehension and production).

5.2 Training & Validation

The training stage of our STGCN involves presenting the model with a time series of observations of childrens' vocabulary changing over time. The hyperpa-

Ground Truth	Head	Torso	Vis	Mouth	Foot	Olf	Gust	Inter	Haptic	Aud	Hand	Nels	Mcrae	Phon	Buch
BAA BAA			•		•		•	•				•		•	•
BABY	•		•					•				•			•
BATH	•	•	•		•	•		•				•	•	•	•
BUBBLES	•		•			•						•			
CHOO CHOO			•												
DADDY	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
GRANDMA	•	•	•	•	•	•	•	•		•	•		•	•	•
GRANDPA	•	•	•	•		•	•					•	•	•	
MEOW		•	•												•
MILK	•	•	•	•	•	•	•	•		•	•	•	•	•	•
MOO			•												
MORE	•		•				•					•	•		•
MUMMY	•		•	•	•	•	•	•		•	•	•	•		•
QUACK	•	•	•			•							•		
WOOF	•	•	•			•					•	•	•	•	•
YES	•	•	•			•		•			•	•	•	•	•
YUM	•	•	•			•		•				•	•	•	•

Table 1. Example prediction made on the same data sample by all models. The Ground Truth column shows new words that have been learned by the child since the previous observation. The dots represent correct prediction of increased knowledge.

rameters of our STGCN were: Epoch size 1000, Batch size 6, optimiser ADAM, the loss function was Mean Absolute Error (MAE), and the metric function was Mean Squared Error. Our Feedforward Neural Network model, trained only on vocabulary data with no relationship element, had hyperparameters: Two hidden layers, Epoch size 1000, Learning Rate 0.8, Batch size 6, Momentum 0.7, Alpha Decay 200, Loss Function Mean Squared Error (MSE).

5.3 Ensemble Models

Given that we were utilising multiple predictive models for comparison in our experiment, an ensemble algorithm was used in order to combine the outputs of the individual models, and potentially improve predictability. There are many approaches to model ensembles [28] and we evaluated the following techniques: *Simple Average*, *Weighted Average*, *Majority Voting*, *OR Classifier* and *AND Classifier*. For each of these techniques, the predictive models' outputs were evaluated to determine an increased level of knowledge, whereby a positive result was obtained when the model predicted productive or receptive knowledge of a word when the most recent observation showed that the child did not possess such knowledge. The *Simple Average* ensemble takes the mean of all individual models to arrive at a final output. The *Weighted Average* obtains the combined output by averaging the individual models with different weights [28], assigning more importance to some models compared to others. We chose to build fifteen Weighted Average ensembles, each giving more weight to a different model, and include the best two performers in the results table. *Majority Voting* obtains a positive result only if more than half of the models have produced a positive prediction. The *OR Classifier* operates in a similar fashion to the OR logic gate, whereby a positive output is obtained if any of the models indicate a positive prediction. Similarly the *AND Classifier* functions like an AND logic gate, producing a positive outcome only when all input models agree on a prediction.

Model	Precision	Recall	F1	Accuracy
Semantic Relationships:				
McRae *et al*	0.32	0.48	0.38	0.58
Buchanan *et al*	0.37	0.41	0.39	0.58
Word Association Relationships:				
Nelson *et al*	0.33	0.41	0.37	0.57
Phonological Relationships:				
BEEP (Jaccard)	0.34	0.53	0.41	0.55
Sensorimotor Relationships:				
Lancaster (Head)	0.30	0.46	0.36	0.53
Lancaster (Gustatory)	0.35	0.35	0.40	0.57
Lancaster (Mouth)	0.32	0.46	0.38	0.59
Lancaster (Olfactory)	0.37	0.47	0.41	0.58
Lancaster (Torso)	0.36	0.42	0.39	0.58
Lancaster (FootLeg)	0.33	0.47	0.38	0.56
Lancaster (Visual)	0.34	0.53	0.41	0.56
Lancaster (Interoceptive)	0.34	0.40	0.37	0.56
Lancaster (Auditory)	0.36	0.46	0.40	0.58
Lancaster (Haptic)	0.39	0.43	0.41	0.59
Lancaster (HandArm)	0.32	0.41	0.36	0.56
Ensembles:				
Simple Average	0.22	0.24	0.23	0.26
Weighted Average (Buch. Semantic)	0.38	0.52	0.42	0.73
Weighted Average (Lanc Haptic)	0.37	0.52	0.43	0.72
'OR' Classifier	0.23	0.81	0.37	0.43
'AND' Classifier	0.06	0.07	0.06	0.36
Majority Vote	0.13	0.36	0.20	0.40
2-Layer Feedforward Neural Network	0.79	0.60	0.68	0.64

Table 2. Results scores of all individual models and ensembles.

5.4 Evaluation

By way of illustration of a typical result, Table 1 displays a randomly selected output from the test dataset for each individual model. Despite similarities in some regions (e.g. all models accurately predicting the appearance of the word 'Daddy' on this observation), it does show some stark differences in predictive accuracy, at least on a per-observation level, as one may expect considering the differences in word relationships.

The 'optimistic' version of the observational data, in which we corrected contradictions in observations by assuming the child did in fact know words that appeared to be 'forgotten', outperformed the 'pessimistic' version, which assumed an observational error by the carer and that the child did not know the word.

Table 2 shows the preliminary results of the fifteen models and six ensembles, plus the output from a Feedforward Neural Network for comparison. In our experiments, the standard Neural Network model displayed the highest accuracy of 0.64, rendering it the best performing individual model. However the Weighted Average ensembles all outperformed the Neural Network, with the Buchanan-emphasised variant performing the best, showing an accuracy of 0.73. The other ensemble algorithms generally showed a decrease in performance.

6 Conclusions and Future Work

We have presented a multi-relationship model that can be used to make predictions about a child's upcoming vocabulary, and the process of constructing it. It has built upon ideas from existing research into infant language acquisition prediction using Neural Networks and graph models, and we have expanded this by considering the current and past vocabularies of a given child combined with multiple relationships between the words. Our findings have shown increased performance of this technique over a standard Neural Network based predictor. Consequently, this technique could serve as a viable foundation for a prospective tool for parents and clinicians, by providing suggestions regarding the most effective words to teach a given child at a particular time for optimal results.

We have identified a number priorities for future development. First, training on more observational data should increase the predictive power of the models. Second, we plan to expand the number of models used to inform the input graphs, including additional psycholinguistic and phonological relationships. This in itself may open up new avenues of research. Third, there may be validity in attempting to optimise the weights used to bias the Weighted Average ensemble. Finally, there are parameters chosen during the process of transforming data from norms into graphs that are worth examining for opportunities to optimise.

References

1. L. Feinstein and K. Duckworth, *Development in the early years: its importance for school performance and adult outcomes*. London: Centre for Research on the Wider Benefits of Learning, 2006.
2. T. S. Scerri *et al.*, "DCDC2, KIAA0319 and CMIP are associated with reading-related traits," *Biol Psychiatry*, vol. 70, no. 3, pp. 237–245, 2011.
3. G. Lindsay *et al.*, "Educational provision for children with specific speech and language difficulties in England and Wales," *IoE and CEDAR*, 2002.
4. J. Clegg *et al.*, "Developmental language disorders - a follow-up in later adult life. cognitive, language and psychosocial outcomes," *J. Child Psychol. Psychiatry*, vol. 46, pp. 128–149, Feb. 2005.
5. S. Roulstone *et al.*, "Investigating the role of language in children's early educational outcomes," Tech. Rep. DFE-RR134, Department of Education, UK, 2011.
6. L. Fenson *et al.*, *MacArthur-Bates Communicative Development Inventories*. Paul H. Brookes Publishing Company Baltimore, MD, 2007.
7. K. J. Alcock *et al.*, "Construction and standardisation of the UK communicative development inventory (UK-CDI), words and gestures," in *International Conference on Infant Studies*, 2016.
8. H. Stadthagen-Gonzalez and C. J. Davis, "The Bristol norms for age of acquisition, imageability, and familiarity," *Behav Res Methods*, vol. 38, no. 4, pp. 598–605, 2006.
9. E. K. Johnson and P. W. Jusczyk, "Word segmentation by 8-month-olds: When speech cues count more than statistics," *J Mem Lang*, vol. 44, pp. 548–567, May 2001.
10. N. Beckage, M. Mozer, and E. Colunga, "Predicting a child's trajectory of lexical acquisition," in *37th Annual Meeting of the Cognitive Science Society, CogSci 2015* (D. C. Noelle *et al.*, eds.), cognitivesciencesociety.org, 2015.

11. N. M. Beckage, M. C. Mozer, and E. Colunga, "Quantifying the role of vocabulary knowledge in predicting future word learning," *IEEE Trans. Cogn. Develop. Syst.*, vol. 12, pp. 148–159, June 2020.
12. K. McRae *et al.*, "Semantic feature production norms for a large set of living and nonliving things," *Behav Res Methods*, vol. 37, pp. 547–559, Nov. 2005.
13. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st Int. Conf. on Learning Representations, ICLR 2013, Workshop Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2013.
14. J. Ke and Y. Yao, "Analysing language development from a network approach," *J Quant Linguist*, vol. 15, no. 1, pp. 70–99, 2008.
15. C. Sims, S. Schilling, and E. Colunga, "Exploring the developmental feedback loop: word learning in neural networks and toddlers," in *Proceedings of the Annual Meeting of the Cognitive Science Society, CogSci 2013* (M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth, eds.), vol. 35, pp. 3408–3413, 2013.
16. D. L. Nelson *et al.*, "The university of South Florida free association, rhyme, and word fragment norms," *Behav Res Methods, Instruments, & Computers*, vol. 36, pp. 402–407, Aug. 2004.
17. L. Buchanan, C. Westbury, and C. Burgess, "Characterizing semantic space: Neighborhood effects in word recognition," *Psychonomic Bulletin & Review*, vol. 8, pp. 531–544, Sept. 2001.
18. T. Robinson, "British English Example Pronunciation (BEEP) dictionary." http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html, 1996.
19. D. Lynott *et al.*, "The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words," *Behav Res Methods*, vol. 52, no. 3, pp. 1271–1291, 2020.
20. M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 2, (Montreal, Que., Canada), pp. 729–734, IEEE, 2005.
21. J. Bruna *et al.*, "Spectral networks and locally connected networks on graphs," *arXiv:1312.6203 [cs]*, May 2014.
22. T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.
23. L. Zhao *et al.*, "T-GCN: A temporal graph convolutional network for traffic prediction," *IEEE trans Intell Transp Syst*, vol. 21, no. 9, pp. 3848–3858, 2020.
24. W. Jiang and J. Luo, "Graph neural network for traffic forecasting: A survey," *Expert Syst Appl*, p. 117921, 2022.
25. CSIRO's Data61, "Stellargraph machine learning library." https://github.com/stellargraph/stellargraph, 2018.
26. M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous systems." https://www.tensorflow.org/, 2015.
27. M. C. Frank, M. Braginsky, D. Yurovsky, *et al.*, "Wordbank: an open repository for developmental vocabulary data," *Journal of Child Language*, vol. 44, pp. 677–694, May 2017.
28. C. Zhang and Y. Ma, eds., *Ensemble Machine Learning*. Springer US, 2012.