Spatio-Temporal Graph Neural Networks for Infant Language Acquisition Prediction

Andrew Roxburgh[0000−0003−1333−5654], Floriana Grasso[0000−0001−8419−6554], and Terry R. Payne[0000−0002−0106−8731]

University of Liverpool, Liverpool, UK
{A.Roxburgh, Floriana, T.R.Payne}@liverpool.ac.uk

Abstract. Predicting the words that a child is going to learn next may be useful for boosting language acquisition, and such predictions have been shown to be possible with both neural network techniques (looking at changes in the vocabulary state over time) and graph model (looking at data pertaining to the relationships between words). However, these models do not fully capture the complexity of the language learning process of an infant when used in isolation. In this paper, we examine how a model of language acquisition for infants and young children can be constructed and adapted for use in a *Spatio-Temporal Graph Convolutional Network (STGCN)*, taking into account the different types of linguistic relationships that occur during child language learning. We introduce a novel approach for predicting child vocabulary acquisition, and evaluate the efficacy of such a model with respect to the different types of linguistic relationships that occur during language acquisition, resulting in insightful observations on model calibration and norm selection.

Keywords: Neural Models of Infants and Child Development · Spatial-Temporal Graph Neural Networks · Language Acquisition Prediction

1 Introduction

Developmental Language Disorder (DLD) is a condition whereby a child shows significant difficulties with language development for no clear reason. It is the most common disability in pre-schoolers, affecting somewhere between 5-7% of all UK children [21, 30, 36]. The condition is exacerbated by other factors associated with learning difficulties, or environmental or familial factors. For example in the UK, between 40% and 56% of children in areas of social disadvantage begin nursery school with such a language delay [19, 22]. This has been proven to cause a knock-on effect within child education; for example, DLD has an incidence rate of 50-90% on reading difficulties [35], which can have a subsequent impact on educational outcomes (in fact reading levels have been shown to be a good predictor of educational outcomes [14]). Furthermore, children show great difficulties catching up with their peers without adequate support [5, 8, 9, 37], and thus are more likely to perform worse academically, to show a lower employment rate, and to have poorer future mental health [7, 32, 42].

Language delay is also an indicator of many other developmental issues; at least 3% of all UK children have a language delay linked with hearing impairment, specific learning difficulties such as dyslexia, and general learning needs [21]. Communication difficulties also feature in Autism Spectrum Disorder, which affects 1% of the UK population [2]. Such challenges are not limited to neurologically 'atypical' children, as the child's communication environment and family circumstances (e.g. poverty or limited language exposure) can have an effect on any early years language development, with a negative, cascading effect on school outcomes years later [26, 38].

With such a public health issue of this scale, identifying affected children early is extremely important. One of the most commonly-used methods for identifying such children is through the use of standardised tools such as the *MacArthur-Bates Communicative Development Inventory (CDI)* [10], a paper instrument originally developed in US English but now adapted into many languages [1]. It consists of a series of questions and checklists designed to assess how children understand and produce words and gestures, as well as their acquisition of grammar. It is typically accompanied by a Family Questionnaire, which gathers information about the family members and individuals who generally spend time with the child during a typical week. This data can be used to generate a *spatial model of language acquisition* (where the spatial component corresponds to words with specific linguistic relationships), which can then be used to identify children who may suffer from language delay.

By automating the process of acquiring this data, through mobile applications such as *Babytalk*[1] [28], a model of a child's language acquisition can be built over time, and used to predict those words that the child should learn in the future [3, 4]. This knowledge can be used as a potential intervention for children with DLD, by suggesting words that a child's carer should emphasise when teaching language, [16], for example, using the *Babytalk* application. This could also be used to boost the language acquisition of neurologically atypical children, and could have further research applications. One way to accomplish this is to refer to *Age-of-Acquisition* norms [33] that can be used to encourage caregivers to emphasise words that typical children would acquire at a similar age. However, children (as individuals) do not tend to learn the same words at the same rate or age, and so a more tailored approach is required. The accuracy of the model used to predict forthcoming words is therefore an important factor.

Graph Neural Networks (GNN) [40] extend the concept of conventional feed-forward neural networks to capture the relationships and interactions between nodes within a graph, by learning a vector representation of each node that not only depends on its own features but also on the features of its neighbours. These *node embeddings* encode the structural and feature data of themselves, their neighbouring nodes, and ultimately extend to the overall graph structure. The resulting model can then be used for edge prediction, node or graph classification, labelling, and feature prediction; or in the context of language acquisition,

[1] http://www.lucid.ac.uk/

it could be used to determine whether the word that the node represents will be 'known' by the child.

In this paper we examine how a model of language acquisition for infants and young children can be constructed and adapted for use in a *Spatio-Temporal Graph Convolutional Network (STGCN)*, taking into account the different types of linguistic relationships that occur during child language learning. We introduce a novel approach for attempting to predict child vocabulary acquisition, and evaluate the efficacy of such a model with respect to the different types of linguistic relationships that occur during language acquisition. Section 2 provides an overview of existing approaches that explore how vocabulary learning can be modelled, and provides a short overview of the STGCN based approach. The way in which a child's vocabulary can be modelled as a graph is presented together with a discussion of how it is used by the STGCN in Section 3. After describing the datasets collected for this study (Section 4), we evaluate the approach empirically (Section 5) and conclude in Section 6.

2 Background

2.1 Previous Approaches to Vocabulary Models

Vocabulary acquisition can be thought of as the evolution of a network of interconnected nodes, where each node models a word and the edges represent a variety of different relationships between them (several of these relationships are discussed in Section 3). As more vocabulary is acquired, the graph expands, and new connections are established. Much of the work in modelling child vocabulary growth has exploited the use of graph representations and artificial neural networks, due to the fact that the words a child learns are inherently connected with each other. Examples include the use of graphs when modelling vocabulary growth over time [17], and the use of neural networks for modelling the way that a brain acquires language [31]. Thus, graph theory can be used as a means of analysing and simulating vocabularies and their evolution.

Beckage, Mozer & Colunga [3] showed that it is possible to predict words that would be learnt in the future using the words a child already knows. By analysing CDI questionnaire data of 77 subjects over a 1-year period at 1-month intervals, and by using a network growth technique, they constructed three different models (*Additive*, *Maximal*, and *Threshold*), each calculating the probability of a word being learnt within the next month in slightly different ways, but all three relying on the probability of a new word being learned as a consequence of the set of words already learnt, and specifically the other words learnt concurrently (i.e. during that same 1-month period). They also successfully developed several neural-network based predictive models using a variety of qualitatively different sources of information as inputs, and observed that the accuracy of predictions could be enhanced by either increasing the temporal resolution of the data, or including more meaningful connections between words in the predictive model.

Subsequent work not only supported the hypothesis that individual words in a child's vocabulary are informative in predicting future vocabulary growth,

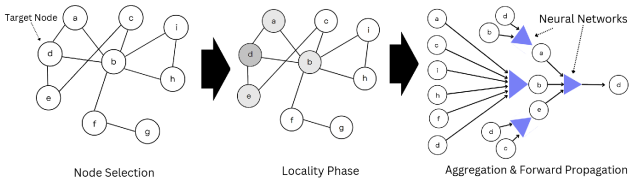

Fig. 1. General operation of a *Spatial Graph Neural Network* (GNN).

but that predictions based on this knowledge reliably outperformed those based purely on demographics [4]. It was also noted that the specific composition of a child's vocabulary significantly affects a model's ability to predict the child's future language acquisition (i.e. the acquisition of new words), suggesting that an individual's vocabulary has relevant and predictive information on the type of learner - and the learning trajectory - of a particular child.

2.2 Overview of Graph Neural Network approaches

Gori [12] and Scarselli [29] introduced *Graph Neural Networks (GNN)* [40], which extend the concept of conventional feedforward neural networks to graph-structured data. This development allowed for more effective processing of this type of data, which historically had posed challenges for many machine learning approaches. Unlike traditional neural networks, which are often designed for handling data in regular structures, such as images or data sequences, GNNs process inherently irregular data with irregular structures, e.g. potentially varying graph sizes, with no fixed number of neighbours for each nodes, and with complex connectivity patterns and attributes. They manage relationships (edges) between entities (nodes/vertices) within a graph, by learning a vector representation of each node (i.e. a *node embedding*) that not only depends on its own features but also on the features of its neighbours. For a vocabulary model, each node in the graph corresponds to a word, and is represented by a feature vector that is initialised with an attribute representing the probability that this word is known.

Spatial GNNs focus on the spatial relationships between nodes, operating directly in the graph's node and edge space. They essentially 'pass messages' between neighbouring nodes, aggregating information from a node's local neighbourhood to learn representations. Fig.1 illustrates the general operation of a Spatial GNN [27]; each node is selected and its local neighbourhood is determined (i.e. the *Locality Phase*), which for a vocabulary model, is governed by the connection between word pairs. The graph structure is then reconfigured (i.e. the *Aggregation Phase*) prior to transforming the aggregated features through a neural network layer resulting in a new feature vector for each node, that is informed by its neighbours (i.e. the *Forward Propagation Phase*). Once all of the nodes have been processed, a final representation (i.e. a node embedding) is obtained for each of the nodes. These embeddings can then be used to determine the probability that the words they represent are 'known'.

Graph Convolutional Networks (GCNs) [20] apply a convolution operation to graphs in GNNs in a similar way to traditional Convolutional Neural Networks

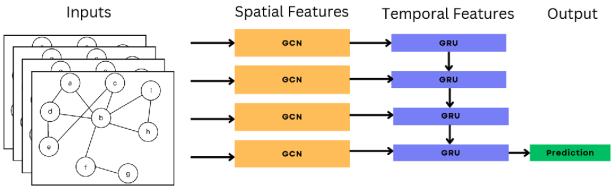

Fig. 2. Block diagram of the T-GCN algorithm [44] used for the Spatio-Temporal Graph Convolutional Network model, using *Graph Convolutional Networks* (GCNs) as spatial features and *Gated recurrent units* (GRUs) as temporal features.

(CNNs), with many filtering methods possible, typically within the Fourier domain (Spectral GCNs) or in the spatial domain (Spatial GCNs). An interesting variation was proposed by Kipf & Welling [18] that merges the two methods, with spectral graph convolutions which are simplified to reduce the overhead that comes with computing a Fourier transform of a graph.

Spatio-Temporal Graph Convolutional Networks (STGCN) further extend GCNs, by combining a time dimension in addition to the spatial dimension, and as such, can be used for handling graph-structured, time-series data. Yan *et al.* [41] first proposed the concept for recognising actions in skeleton models, and similar designs have been described as a method of predicting traffic in road networks [43, 44], with the data coming from road networks and traffic sensors represented as a graph, where edges represents direct routes between sensors, and the time dimension referring to the history of the sensor data [15].

STGCN algorithms, such as T-GCN [44] shown in Fig. 2, typically employ a hybrid design that combines a graph convolutional network block, with a Recurrent Neural Network block such as a *gated recurrent unit* (GRU), designed for sequential data. The graph convolutional network is used to capture the relationships between nodes, thus modelling spatial dependence, and the gated recurrent unit is used to capture the dynamics of the node features, modelling temporal dependence. STGCNs allow the modelling of evolving graph structures in which node dependencies change over time, which makes them a promising tool for attempting to predict a child's future vocabulary on the basis of a dynamic vocabulary together with fixed relationships between words.

3 Modelling Vocabulary as a Graph

In a basic representation of the child vocabulary as a graph, each node represents a word with an associated feature vector, incorporating information about the state of the word. We can represent the depth of the child's knowledge of each word by including a feature quantifying it. Each feature is initialised with discrete values that indicate that a child may: i) understand a word but cannot produce it; ii) produce a word but does not understand it; iii) both understand and produce a word; or iv) have no knowledge of a word. Whilst this model may be a possible oversimplification of a child's word knowledge and cognition, these discrete states have the advantage of being easily observable and well understood.

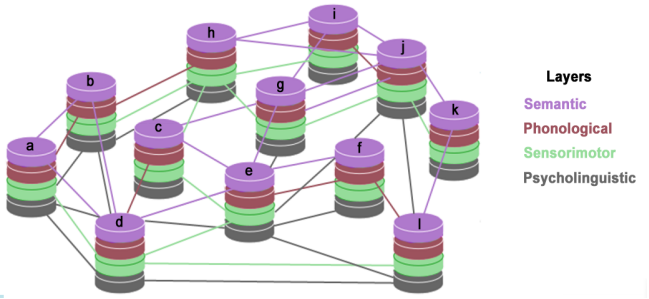

Fig. 3. Simplified example of a *Multiplex Lexical Network* with 4 relationship layers and 12 words represented by letters `a` - `l`. Edges are weighted (but not shown here).

The model can be enhanced by incorporating quantified relationships between words (thereby defining and weighting the graph edges). The various graphs can be arranged as a *multiplex* [39] graph (Fig. 3), which is a type of graph where nodes represent the same set of items on all layers. In the context of language, this has also been described as a *Multiplex Lexical Network (MLN)* [34]. A variety of data exists that could be used for this purpose, including:

Word Associations: words that have some cognitive association with each other - for example `lock`/`key`, or `umbrella`/`rain`;
Phonological Connections: data on words that that have a measurable similarity due to sharing similar phonemes;
Psycholinguistic Norms: how easy a word can be mentally visualised (i.e. *Imageability*);
Semantic Features: connecting words of a similar meaning in context;
Concretedness: whether a word represent a concrete or an abstract concept;
Familiariy: how commonly used a word is; or
Other: simple meausres such as how difficult a word is to remember or pronounce (e.g. word length).

Whilst a number of relationships exist that could be included, care needs to be taken not to rely too heavily on an 'adult perspective', but rather concentrate on those more in line with an infant's cognitive perspective. In this paper, we discuss the incorporation of the following set of norms in the model:

Semantic Feature Production norms: these measure the semantic similarity of words. We focus on using the McRae [24] and Buchanan [6] norms.
Sensorimotor norms: these measure perception and action strength of words from the perspective of the child. We use the *Lancaster Sensorimotor Norms* [23], which evaluate English words based on six perceptual modalities (i.e., *touch*, *hearing*, *smell*, *taste*, *vision*, and *interoception*) and five action effectors (i.e., *mouth/throat*, *hand/arm*, *foot/leg*, *head excluding mouth/throat*, and *torso*).

Sensorimotor norms are named after the initial phase in Piaget's theory of cognitive development [25], which spans from birth to approximately 2 years of age.

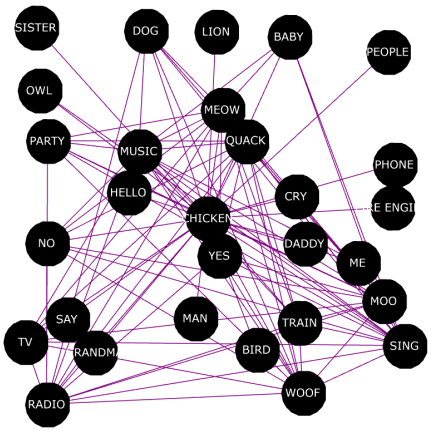

Fig. 4. Example *auditory sensorimotor* relationship graph (simplified) focused on the word `chicken` (based on Lancaster Sensorimotor Norms).

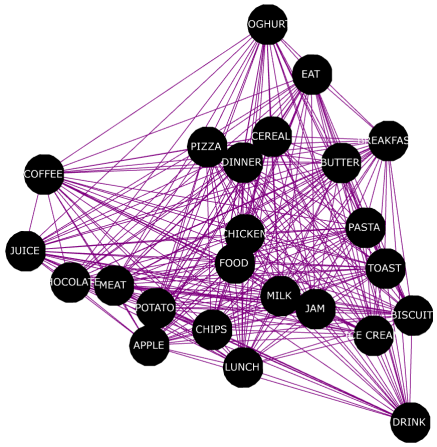

Fig. 5. Example *gustatory sensorimotor* relationship graph (simplified) focused on the word `chicken` (based on Lancaster Sensorimotor Norms).

During this period, infants learn to use their senses to construct an understanding of the world, and use motor movements (*reaching*, *sucking*, *grasping*, and *touching*) to interact with it. Thus, the primary objective of the sensorimotor stage is to develop an understanding of *object permanence*, i.e. the realisation that objects and events persist even when they are not directly observable by the child. From the perspective of infant language acquisition, it is especially useful to be able to connect words from the conceptual point of view of a child at the earliest development stage. Figs. 4 and 5 illustrate a simplified vocabulary graph with edges connecting nodes with similar *Lancaster Norm* scores: Fig. 4 illustrates nodes in an auditory graph; whereas Fig. 5 illustrates those in a gustatory one. Both graphs are simplified and focused on the word `chicken`.

To represent the strengths of connections between words in the graph for each of the similarity measures we used, in the case of the Semantic Feature Production Norms, the necessary cosine similarity matrices were used together with the data by McRae and Buchanan. Lancaster norms require additional processing for use in our model, and thus a weighted adjacency matrix was created for all possible word pairs within each category by computing the cosine similarity of each pair, which in this case is simply the normalised product of their scores. Thus, words with, for example, high scores in the *Smell* category (e.g. `orange` and `lemon`) would have a strong connection, whilst a word with high score and a word with low score (e.g. `orange` and `table`) would have a weak connection. Similarly, the connection between two words with low scores in the *Smell* category (e.g. `table` and `chair`) would also be weak. In addition, all self-loops (e.g. `head` and `head`) have a maximal weight (i.e. 1.0).

For each category, the resulting matrix comprised approximately 76,000 data points. Given that many of the pairs may not be relevant to the category and

would result in an unnecessary computational burden, a threshold was used (with the inter-word connection value of 0.5) to eliminate pairs deemed unnecessary due to the implied lack of connection. For each sensorimotor category, we used the adjacency matrices to define the edges, and created a list of nodes by removing duplicates in the edge list. Finally, a feature was also added to each node representing the level of knowledge of that word.

4 Data Collection

Two types of data were collected to build a model and evaluate it; datasets of words and their relationships were collected in order to create the model itself (i.e. the *Relational Datasets*), and observational data on child vocabulary was collected in order to test the resulting model (i.e. the *Observational Dataset*). For the *Relational Datasets*, we used the English language relationship datasets comprising both semantic and sensorimotor norms from [6, 23, 24]. A challenge in putting together independent datasets into one model is one of standardisation. This should be addressed by accounting for differences due to: i) synonyms (e.g. `rabbit` and `bunny`); ii) regional variations (e.g. `mommy` and `mummy`); and iii) international spelling variations (e.g. `colour` and `color`). This was addressed by pre-processing the data to identify such cases, and renaming words for consistency, thus resulting in a *standardised* vocabulary.

A second challenge was to consider *homographs*, i.e. words with multiple meanings such as `back`, `bank`, `run`, or `drink`. This was address by using a contextual specifier as a label for ambiguous words, thus ensuring that the vocabulary was unambiguous. For example, ambiguous words such as `drink` become `drink(beverage)` and `drink(verb)`. This is particularly challenging when considering the child perspective, as for certain words, such relationships are different than for adults. To illustrate this, the notion 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 simultaneously both a food and an animal. Again these words were appended with a context-appropriate label.

The *Observational Dataset* consisted of data collected through an experiment with participants using our web app as part of a larger study[2] as well as item-level CDI Survey responses extracted from all available forms in English acquired from Wordbank [11]. We selected only data for which there are longitudinal sequences, to enable the T-GCN algorithm's GRU layer block to train on temporal data. Words were converted to our standardised vocabulary to allow for Irish, Australian, American and regional British dialect differences. The overall dataset included 460 observations (i.e. vocabulary inventories), in sequences of 4 contiguous observations, using a sliding window method to create smaller sequences from longer ones. The training data consisted of 115 'test subjects' corresponding to individual sequences (as a child may generate more

[2] The study has received ethical approval by the Faculty of Science and Engineering Research Ethics Committee, University of Liverpool, Ref. 8182.

than one sequence), and the test dataset consisted of 120 observations across 30 'test subjects'.

Due to the nature of real-world, human-collected data regarding human behaviour, inevitably there will be errors present, and thus working with such data presents challenges. With a large dataset containing millions of observations, occasional errors will have minimal impact on the overall behaviour. However, such errors may have an adverse effect on trials with smaller datasets (such as that used for this evaluation), and thus, to address this, the data was pre-processed to identify possible errors and eliminate discrepancies. Common errors encountered in observational data include:

Observer Misinterpretation: a child may be observed at one time period as understanding, but not producing, a particular word, however it becomes apparent that the child has a different interpretation for that word.

Mistaken Observation: The parent believes that a child has said and/or understood a word, but they change their mind in the subsequent observation.

Observational Error: a sequential pair of observations may have made sense to the parent at the time of each observation, but as a sequence they are confusing, because they suggest that the child has learned a word and then either forgotten the meaning of the word, or forgotten how to say it. For the purposes of our experiments, we assume that children do not forget words once they have demonstrated knowledge of them.

Unobserved knowledge: The parent may not be aware that a child knows a particular word. The child may have used the word at a grandparents' house for instance.

Articulation Error: The child may speak a certain word, but pronouncing it in such a way that it is unintelligible or ambiguous to an observer.

Dropping words from the observed vocabulary could adversely affect the ability to train a classifier on the data; therefore to deal with contradictory data, two datasets were generated, an *optimistic* one whereby it was assumed that the child continue to understand the word in subsequent observation periods, and a *pessimistic* one whereby the assumption is that words are false observations, and that the child did not understand the word during the first observation.

5 Implementation and Evaluation

A collection of thirteen ($N = 13$) individual models were used for the vocabulary relationships explored in this study, represented as graphs $G_N = (V_N, E_N)$ that were based on each of the norms described in the data model (Section 3), where V_i is the set of vertices or nodes, and E_i are the corresponding set of edges for each model $i \in N$. A set of nodes was created for every observation in the data, and was populated with the corresponding observed data. These sets of nodes were combined to form a time series. The set of edges were then processed by combining each edge set with each node set in the time series, resulting in the creation of a time series of graphs for each of the vocabulary relationships. The aim

Table 1. Experimental HyperParameters used by all STGCN models

Parameter	Value	Parameter	Value
Data Mode	*Optimistic*	Epochs	*1000*
Batch Size	*4*	LSTM Sequence Length	*4*
Prediction Length	*1*	Loss Optimiser	*ADAM*
Loss Metric	*MAE*	Model Metric	*MSE*
Dropout	*None*	Input Graph Connections Limit	*2000*

of the model is to predict a child's future vocabulary based on past and existing knowledge, represented as a series of graphs representing different relationships between words in a vocabulary over a number of discrete time periods. Therefore, the nodes were embedded with feature vectors representing the child's current knowledge of the word at each time period. 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: i) no comprehension (0.0); ii) production without understanding (0.3); iii) understanding but no production (0.6); and iv) full comprehension and production (1.0).

A Spatio-temporal Graph Convolutional Network-based model was developed in Python using the *Stellargraph*[3] software library, and configured using the hyperparameters in Table 1. Stellargraph is built on *Tensorflow*,[4] and facilitates the construction of graph-based machine learning models. The Graph Neural Network algorithm used in this study is based on the Temporal-GCN (T-GCN) as described by Zhao et al. [44]. The full model consists of 13 relationship layers, each of which is a separate STGCN model that has been individually trained and executed. Some nodes in these layers may be unconnected as they have no meaningful associations with other words in the context of that relationship.

When a new prediction is required, the vector representing the child's current vocabulary is used to populate the feature vectors of each node on each relationship layer, representing the new words that the child had recently learned. The GCN classifier, in conjunction with the STGCN's spatio-temporal block, is then applied to these graphs to re-classify the unknown nodes. These newly classified nodes can then be used to determine the words that are likely to be influenced the most by its neighbours, and so may be learned next. This results in a list of candidate words from each relationship layer, from which the most likely words a child should learn can be determined.

5.1 Training & Validation

The training stage of our STGCN involves presenting the model with a time series of graphs - analogous to a spatio-temporal graph - representing observations of childrens' vocabulary changing over time. This consists of a chain of 4-consecutive-observation chunks. The Sequence Length parameter of the STGCN

[3] https://github.com/stellargraph/stellargraph
[4] https://www.tensorflow.org/

Table 2. Results for each of the individual models.

Reference (baseline) Model	Precision	Recall	F1	Accuracy
2-Layer Feedforward (ANN) Neural Network	0.283	0.854	0.426	0.610
Semantic Relationships:				
McRae *et al.*	0.450	0.513	0.479	0.740
Buchanan *et al.*	0.403	0.606	0.484	0.715
Sensorimotor Relationships:				
Lancaster (Head)	0.443	0.509	0.474	0.737
Lancaster (Gustatory)	0.427	0.598	0.498	0.731
Lancaster (Mouth)	0.417	0.637	0.504	0.716
Lancaster (Olfactory)	0.424	0.571	0.487	0.733
Lancaster (Torso)	0.438	0.537	0.483	0.741
Lancaster (Foot/Leg)	0.428	0.548	0.480	0.733
Lancaster (Visual)	0.438	0.618	0.513	0.732
Lancaster (Interoceptive)	0.435	0.494	0.462	0.739
Lancaster (Auditory)	0.465	0.395	0.427	0.750
Lancaster (Haptic)	0.419	0.586	0.488	0.730
Lancaster (HandArm)	0.404	0.540	0.462	0.722

was set to match the consecutive observation chunk size. We split the input data into training, testing and validation sets.

A *2-layer Feedforward Neural Network* model was also trained using vocabulary data with no relationship element, to provide a baseline for the evaluation. Each word in the child's vocabulary was linked to an input node in the network, which connects to a hidden layer with 500 units. Each word is represented as a node in the output layer. Other model hyperparameters include a Learning Rate of 0.8 and Momentum of 0.9, and the network was trained over 1000 epochs.

5.2 Evaluation

A total of thirteen models were evaluated (2 that model semantic relationships and 11 that model sensorimotor relationships, based on the norms described in Section 3), and compared with the baseline 2-layer Feedforward Neural Network. The aim is to evaluate each of these models independently, and compare them with the simple neural network approach. The results are presented in Table 2, together with the following evaluation metrics:

Precision: Measures the accuracy of positive predictions; i.e. out of all the positive predictions (i.e. words that the model has predicted will have increased comprehension), how many of those were proven to be correct.

Recall: Measures the fraction of relevant instances retrieved; i.e. out of all the words that were actually proven to be learned, how many were correctly predicted by the model.

F1 Score (or F1 Measure): Harmonic mean of precision and recall.
Accuracy: Measures the fraction of correct predictions, representing the ratio of correctly predicted words with increased comprehension, to the total number of new words actually learned by the child.

Although the overall performance of the ANN demonstrated moderate performance, it fell short of that demonstrated by each of the STGCN models (Table 2). The results suggests that although many of the words were predicted, the percentage of those being correct was low. In comparison, the use of the different relationships with the STGCN model resulted in a greater accuracy of the words predicted, but with a drop in the overall percentage of possible predictions.

The accuracy of the individual predictive models varied greatly despite achieving similar overall accuracy across all models for the entire test set. There is some variance in the performance of the different STGCN models incorporating Sensorimotor relationships, with *Visual* and *Mouth* models showing notably higher recall values at 0.618 and 0.637 respectively. These high recall values are indicative of the models' ability to identify a larger proportion of relevant words for learning, making them suitable choices for minimising false negatives. The *Auditory* STGCN model clearly stands out with the highest accuracy (0.750) and highest f1 (0.465) but the lowest recall (0.395), suggesting it is more selective but might miss out on some words the child could learn. The *Haptic*, *Gustatory*, and *Olfactory* models balance between f1 and recall, with f1-scores in the 0.488 to 0.498 range, indicating a reasonable compromise between the metrics. Overall, f1 for the Sensorimotor STGCN models is relatively high, ranging from 0.716 to 0.750, suggesting they make correct predictions about whether a word will or will not be learned most of the time. The *Torso*, *Foot/Leg*, and *Interoceptive* models show neither the highest nor the lowest performance but could still potentially provide useful predictions for vocabulary acquisition The performance when using the semantic relationships was comparable to those of the sensorimotor relationships, with the McRae et al. norms resulting a model that performed marginally better than that of the Buchannan et al. norms.

A comparative evaluation was also conducted to compare the two variants of the observational data: *optimistic* and *pessimistic*, with the former correcting contradictions in observations by assuming the child knew the words that appeared to be forgotten, while the latter assumed an observational error by the caregiver and that the child did not know the word. In general, the *optimistic* version outperformed the *pessimistic* version, thus indicating the significance of the observational data in determining the accuracy of predictive models.

6 Conclusions & Discussion

A Spatio-Temporal Graph Neural Network model (and the way in which it can be constructed) was presented that can be used to make predictions about a child's upcoming vocabulary acquisition. The model is based on existing work on infant language acquisition prediction using neural networks and graph models; however, we have expanded this by considering the current and past vocabularies

of a given child combined with multiple types of relationships between the words, including sensorimotor norms and semantic feature prediction norms.

The overall accuracies of each model were reasonably similar; however analysis of outputs at the individual prediction level showed several differences. This suggests that there could be scope for exploitation of the variation of coverage. One recognised way to combine and improve the predictability of multiple models is to explore the use an ensemble output stage [13]. This would allow each of the models to contribute towards the overall prediction, but would involve an investigation into different approaches for generating ensembles (such as Boosting, Bagging, Stacking, Voting and Averaging). A longer term plan is to incorporate more observational data into our model inputs, which may increase the predictive power of the models. We also aim to expand the number of relationships we use to inform the input graphs, including word association norms, phonological relationships, and psycholinguistic vocabulary norms such as imageability and concreteness. Furthermore, assigning a feature to each node in the graph that represents the length of time that the word has been understood by the child, such that words understood for longer carry more influence (or even vice versa) would also be worthy of investigation.

References

1. Alcock, K.J., Rowland, C.F., Meints, K., Christopher, A.E., Just, J., Brelsford, V., Summers, J.: Construction and standardisation of the UK communicative development inventory (UK-CDI), words and gestures. In: International Conference on Infant Studies (2016)
2. Baron-Cohen, S., Scott, F.J., Allison, C., et al.: Prevalence of autism-spectrum conditions: UK school-based population study. Br J Psychiatry **194**(6), 500–509 (Jun 2009)
3. Beckage, N., Mozer, M., Colunga, E.: Predicting a child's trajectory of lexical acquisition. In: Noelle, D.C., et al. (eds.) 37th Annual Meeting of the Cognitive Science Society, CogSci 2015 (2015)
4. Beckage, N.M., Mozer, M.C., Colunga, E.: Quantifying the role of vocabulary knowledge in predicting future word learning. IEEE Trans. Cogn. Develop. Syst. **12**(2), 148–159 (Jun 2020)
5. Bleses, D., Makransky, G., Dale, P.S., et al.: Early productive vocabulary predicts academic achievement 10 years later. Appl Psycholinguist **37**(6), 1461–1476 (Nov 2016)
6. Buchanan, E.M., Valentine, K.D., Maxwell, N.P.: English semantic feature production norms: An extended database of 4436 concepts. Behav Res Methods **51**(4), 1849–1863 (Aug 2019)
7. Clegg, J., Hollis, C., Mawhood, L., et al.: Developmental language disorders - a follow-up in later adult life. cognitive, language and psychosocial outcomes. J. Child Psychol. Psychiatry **46**(2), 128–149 (Feb 2005)
8. Conti-Ramsden, N., Botting, Z., Simkin, E., et al.: Follow-up of children attending infant language units: Outcomes at 11 years of age. Int J Lang Commun Disord **36**(2), 207–219 (2001)

9. Feinstein, L., Duckworth, K.: Development in the early years: its importance for school performance and adult outcomes. Tech. rep., Centre for Research on the Wider Benefits of Learning, London (2006), https://discovery.ucl.ac.uk/id/eprint/10005970/1/Feinstein2006Development.pdf
10. Fenson, L., Marchman, V., Thal, D., Dale, P., Reznick, J., Bates, E.: MacArthur-Bates Communicative Development Inventories. Paul H. Brookes Publishing Company, Baltimore, MD, 2nd edn. (2007)
11. Frank, M.C., Braginsky, M., Yurovsky, D., et al.: Wordbank: an open repository for developmental vocabulary data. Journal of Child Language **44**(3), 677–694 (2017)
12. Gori, M., Monfardini, G., Scarselli, F.: A new model for learning in graph domains. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. vol. 2, pp. 729–734. IEEE, Montreal, Que., Canada (2005)
13. Hansen, L., Salamon, P.: Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence **12**(10), 993–1001 (1990). https://doi.org/10.1109/34.58871
14. Hulme, C., Snowling, M.J.: Developmental disorders of language learning and cognition. John Wiley & Sons, Chichester (2013)
15. Jiang, W., Luo, J.: Graph neural network for traffic forecasting: A survey. Expert Systems with Applications **207**, 117921 (2022)
16. Johnson, E.K., Jusczyk, P.W.: Word segmentation by 8-month-olds: When speech cues count more than statistics. J Mem Lang **44**(4), 548–567 (May 2001)
17. Ke, J., Yao, Y.: Analysing language development from a network approach∗. Journal of Quantitative Linguistics **15**(1), 70–99 (Feb 2008)
18. Kipf, T., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations, ICLR 2017 (Poster) (2017). https://doi.org/https://doi.org/10.48550/arXiv.1609.02907
19. Law, J., McBean, K., Rush, R.: Communication skills in a population of primary school-aged children raised in an area of pronounced social disadvantage. Int J Lang Commun Disord **46**(6), 657–64 (2011)
20. Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J.: A survey of convolutional neural networks: Analysis, applications, and prospects. IEEE Transactions on Neural Networks and Learning Systems **33**(12), 6999–7019 (2022)
21. Lindsay, G., Dockrell, J., Mackie, C., et al.: Educational provision for children with specific speech and language difficulties in England and Wales. Tech. rep., CEDAR, University of Warwick (2002)
22. Locke, A., Ginsborg, J., Peers, I.: Development and disadvantage: implications for the early years and beyond. International Journal of Language and Communication Disorders **37**(1), 3–15 (Jan 2002)
23. Lynott, D., Connell, L., Brysbaert, M., et al.: The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. Behav Res Methods **52**(3), 1271–1291 (2020)
24. McRae, K., Cree, G.S., Seidenberg, M.S., et al.: Semantic feature production norms for a large set of living and nonliving things. Behavior Research Methods **37**(4), 547–559 (Nov 2005)
25. Piaget, J.: The Origins Of Intelligence In Children. Routledge, London (1953)
26. Roulstone, S., Law, J., Rush, R., et al.: Investigating the role of language in children's early educational outcomes. Tech. Rep. DFE-RR134, Department of Education, UK (2011)
27. Roxburgh, A.: Using Graph Neural Networks to Predict Toddler Vocabulary Acquisition. Ph.D. thesis, University of Liverpool, Liverpool, UK (2024)

28. Roxburgh, A., Grasso, F., Payne, T.: Predicting word learning to boost child language acquisition. 7th International Conference on Digital Health, DH '17, London, UK (poster session) (July 2017)
29. Scarselli, F., Sweah Liang Yong, Gori, M., et al.: Graph neural networks for ranking web pages. In: The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05). pp. 666–672. IEEE (2005)
30. Scerri, T.S., Morris, A.P., Buckingham, L.L., Newbury, D.F., Miller, L.L., Monaco, A.P., Bishop, D.V., Paracchini, S.: DCDC2, KIAA0319 and CMIP are associated with reading-related traits. Biol Psychiatry **70**(3), 237–245 (2011)
31. Sims, C., Schilling, S., Colunga, E.: 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. vol. 35, pp. 3408–3413 (2013)
32. Snowling, M.J., Adams, J.W., Bishop, D.V., et al.: Educational attainments of school leavers with a preschool history of speech-language impairments. International Journal of Language and Communication Disorders **36**(2), 173–183 (2001)
33. Stadthagen-Gonzalez, H., Davis, C.J.: The Bristol norms for age of acquisition, imageability, and familiarity. Behav Res Methods **38**(4), 598–605 (2006)
34. Stella, M., Beckage, N.M., Brede, M.: Multiplex lexical networks reveal patterns in early word acquisition in children. Scientific Reports **7**(March), 1–10 (2017). https://doi.org/10.1038/srep46730
35. Stothard, S.E., Snowling, M.J., Bishop, D.V.M., et al.: Language-impaired preschoolers. Journal of Speech Language and Hearing Disorders **41**(2), 407–418 (1998)
36. Tomblin, J.B., Records, N.L., Buckwalter, P., Zhang, X., Smith, E., O'Brien, M.: Prevalence of specific language impairment in kindergarten children. Journal of Speech Language and Hearing Disorders **40**(6), 1245–1260 (Dec 1997)
37. Van Dulm, O., Southwood, F.: Does socioeconomic level have an effect on school-age language skills in a developed country? Stellenbosch Papers In Linguistics Plus **49** (Dec 2016)
38. Walker, D., Greenwood, C., Hart, B., et al.: Prediction of school outcomes based on early language production and socioeconomic factors. Child Development **65**(2), 606 (Apr 1994)
39. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Structural Analysis in the Social Sciences, Cambridge University Press, Cambridge (1994). https://doi.org/10.1017/CBO9780511815478
40. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S.: A comprehensive survey on Graph Neural Networks. IEEE Transactions on Neural Networks and Learning Systems **32**(1), 4–24 (2021)
41. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on Artificial Intelligence. vol. 32(1), pp. 7444–7452 (2018)
42. Young, A.R., Beitchman, J.H., Johnson, C., et al.: Young adult academic outcomes in a longitudinal sample of early identified language impaired and control children. Journal of Child Psychology and Psychiatry **43**(5), 635–645 (Jul 2002)
43. Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI) (jul 2018)
44. Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., Deng, M., Li, H.: T-GCN: A temporal graph convolutional network for traffic prediction. IEEE transactions on intelligent transportation systems **21**(9), 3848–3858 (2020)