Towards Using Diachronic Distributed Word Representations as Models of Lexical Development

Anonymous ACL submission

Abstract

Recent work has shown that distributed word representations can encode abstract information from child-directed speech. In this paper, we use diachronic distributed word representations to perform temporal modeling and analysis of lexical development in children. Unlike all previous work, we use temporally sliced corpus to learn distributed word representations of child-speech and child-directed speech under a curriculum-learning setting. In our experiments, we perform a lexical categorization task to plot the semantic and syntactic knowledge acquisition trajectories in children. Next, we perform linear mixed-effects modeling over the diachronic representational changes to study the role of input word frequencies in the rate of word acquisition in children. We also perform a fine-grained analysis of lexical knowledge transfer from adults to children using Representational Similarity Analysis. Finally, we perform a qualitative analysis of the diachronic representations from our model, which reveals the grounding and word associations in the mental lexicon of children. Our experiments demonstrate the ease of usage and effectiveness of diachronic distributed word representations in modeling lexical development.

1 Introduction

Human-like linguistic generalization plays a key role in developing better models for natural language processing (Linzen, 2020). Modeling the lexical development in children is an important aspect of demystifying the dynamics of human language learning. Lexical development in children is a holistic and complex phenomenon involving noisy multimodal interactions and underlying various psycholinguistic processes. Previous research in child language acquisition has shown that infants are capable of lexical processing of words through their semantic and syntactic distributional structures (Lany and Saffran, 2010; Syrett and Lidz, 2010). Recently, the paradigm of word embeddings from deep-learning-based computational semantics has pushed the frontiers in modeling such distributional structures of words (Mikolov et al., 2013a; Wang et al., 2020). Consequently, word embeddings have been used to study various aspects of child-speech and child-directed adult speech (Huebner and Willits, 2018b; Fourtassi et al., 2019; Fourtassi, 2020).

Recent advances in computational modeling for distributional semantics have made it possible to study the diachronic semantic shifts in a given corpus (Kutuzov et al., 2018). Using temporally-wide large-scale corpora, diachronic word embeddings can be used to study the underlying linguistic and non-linguistic dynamics of change and development in human language (Kutuzov et al., 2018). Given the availability of such corpora for child-speech and child-directed adult speech (MacWhinney, 2000), a similar framework can be designed to model the lexical development in children.

This paper explores the usability of diachronic distributed word representations[1] in cognitive modeling and analysis of the lexical development in children. Unlike previous work, we use temporally sliced data to learn distributed word representations of child-speech and child-directed speech under a curriculum-learning-like setting. Through our experiments we show that diachronic word representations can be very effective in capturing various empirical and qualitative aspects of lexical development in children.

2 Background

The meanings of words change over time, owing to a variety of linguistic and non-linguistic factors. This phenomenon has been termed as semantic shifts (Bloomfield, 1933) in historical linguistics.

[1] the terms "distributed word representations", "word vectors", and "word embeddings" have been used interchangeably in this paper.

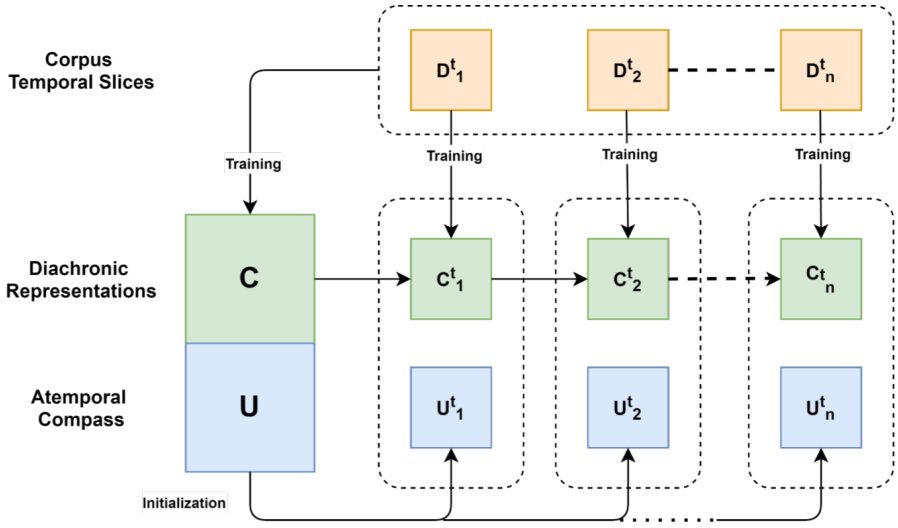


Figure 1: The architecture for the incremental diachronic compass-based word embedding model. In order to temporally align the word embeddings from each of the corpus temporal slices, one of the layers (U_i^t) is initialized from the atemporal compass and frozen during the temporal fine-tuning. The representations for each time-step C_i^t are initialized with the fine-tuned representations for the previous time-step C_{i-1}^t, and fine-tuned on its corresponding corpus temporal slice D_i^t (Section 3.2).

Due to an increasing availability of large corpora, several data-driven methods have been proposed to study such semantic shifts. Words can be represented as continuous vectors (Rumelhart and McClelland, 1986; Elman, 1990). As the meaning and context of a word changes, it's vector changes accordingly (Kutuzov et al., 2018). Recent work has shown that semantic shifts are not always closely related to the changes in word frequencies (Kutuzov et al., 2018). Previous works have also shown that distributed word representations (Turney and Pantel, 2010; Baroni et al., 2014) outperform frequency based methods in detecting semantic shifts (Kulkarni et al., 2014). These models use deep-learning based word-embedding vectors (Mikolov et al., 2013b), produced from word co-occurrence relationships.

Semantic shifts were first analyzed on year-wise data by Kim et al. (2014) using a prediction-based word embedding model. Kim et al. (2014) used a distributional continuous skip-gram model with negative sampling (Mikolov et al., 2013b), which was later proven to be superior in semantic shift analysis as compared to PPMI-based distributional models (Hamilton et al., 2016). Most modern word embedding models are inherently stochastic (Kutuzov et al., 2018), and produce word representations in different vector spaces on each run. To overcome this, the models need to be aligned to one common vector space. This can be done by performing certain linear transformations on the diachronic word embeddings (Kulkarni et al., 2014; Zhang et al., 2015). More recently, Carlo et al. (2019) proposed a relatively simple temporal compass based alignment, which showed significant improvements over the previous approaches.

Diachronic embeddings have been used across a variety of applications, ranging from linguistic studies to cultural studies. They have been successfully applied to tasks like event detection, predicting civil turmoils, and tracing popularity of entities (Kutuzov et al., 2018). Even though there have been a significant number of studies which use static distributed word representations for analyzing child-directed speech (Huebner and Willits, 2018a,b; Fourtassi et al., 2019; Fourtassi, 2020), the usability of diachronic embeddings as models of lexical development has not been explored yet.[2] The closest work is presented by Huebner and Willits (2018a), where they train sequential deep-learning models on age-ordered child-directed speech data. In this study, we use diachronic word embeddings trained on child-speech data,[3] to construct a temporal representational model for the mental lexicon in children.

[2] A parallel study by Jiang et al. (2020) focuses on using diachronic word embeddings to study the child-directed speech. Whereas, in this work we focus on directly modeling the lexical development in children.

[3] This has not been explored before to the best of our knowledge

3 Modeling

Given a corpus of child-speech, a diachronic word-embedding model can be trained over its temporal slices. The distributed-latent representations from the model can then be probed for lexical knowledge at any given point of time. Consequently, the *lexical development* can be simply captured by comparing these distributed representations over some interval of time. In this section, we describe our cognitively motivated diachronic modeling method (Section 3.2), and the required pre-processing of the child-speech corpus (Section 3.1). We discuss the usability of the trained diachronic distributed representations for modeling lexical development in Section 4.

3.1 Data

Similar to all the previous works, we use the CHILDES corpus (MacWhinney, 2000) for our experiments. The corpus consists of speech-transcripts of first language acquisition by children. It contains transcriptions in 26 languages, spanning across 130 corpora of children interacting in different environments, including spontaneous interactions, as well as controlled classroom learning. For the current study, we use the data from the American-English speaking children. Due to an imbalance in the data distribution, we discard the data beyond the first three years of age.

Unlike child-directed adult speech, child speech is highly noisy in the early months. It is only after the age of 18 months, that children start combining two words or single-word phrases in situations in which they both are relevant and having roughly equivalent status (Bavin and Naigles, 2017). Hence, we consider the data after the age of 18 months only. This results in a corpus that is temporally spread across 19 months (age=18 months to age=36 months). It contains 2798 speech transcripts; $1,321,772$ word tokens; $13,812$ word types; and $405,596$ utterances, collected from 28 different studies involving 188 children (99F and 89M) and their guardians.

We tokenize the corpus with whitespace delimitation. We remove all the punctuation from the corpus as they do not contribute lexically in any way. Further, all the proper nouns are replaced with a generalized token: $[NAME]$. We set the temporal granularity of our diachronic models to a month. Hence, we split the corpus into 19 month-wise temporal slices.

3.2 Model

A diachronic word embedding model usually comprises two major components: a base word-embedding model, and a mechanism to align the representations across different temporal data slices. We use $word2vec$ as our base word-embedding model (skip-gram with negative sampling - SGNS variant) (Mikolov et al., 2013b). For aligning the word-embeddings, we employ a slightly modified version of the compass-based alignment method proposed by Carlo et al. (2019).

$word2vec$ is a shallow, two-layered neural-network word-embedding model. It takes a large corpus of words as input, and generates a multi-dimensional distributed vector space, with each word being assigned a vector. The $word2vec - SGNS$ model in particular takes a one-hot encoded word identity vector as input and predicts its surrounding context words. Formally, given a sequence of input words $w_1, w_2, w_3....w_T$, the objective of the model is to maximize the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c<j<c, j\neq 0} logp(w_{t+j}|w_t) \qquad (1)$$

where $c = 5$ is the size of the training context window. The skip-gram formula defines $p(w_{t+j}|w_t)$ using a negative sampling method (Mikolov et al., 2013b).

Some words are quite frequent in the corpus while others are less frequent; to deal with this, a sub-sampling strategy is used. Each word in the training set w_i is discarded with the probability:

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}} \qquad (2)$$

where $f(w_i)$ is the frequency of the word and t is some selected threshold. This accelerates the learning process and improves the accuracy of the vector representations of rare words. We use a word representation dimension of $d = 100$ for the $word2vec - SGNS$ model.

To work with diachronic word embeddings, the embedding spaces generated by the models trained at each time-step need to be aligned to a common embedding space. This ensures that the embeddings across different time-steps can be compared directly. We implement a temporal compass-based model to align the embedding spaces (Carlo et al.,

2019). As the $word2vec - SGNS$ model is a two-layered neural network, the context embeddings are encoded in the output layer parameters (U) and target word embeddings are encoded in the input layer parameters (C) (Figure 1). Given a corpus D, divided into n temporal slices: $D_1^t, D_2^t, ...D_n^t$; the model (C, U) is first trained on the entire corpus $D = D_1^t, D_2^t, ...D_n^t$ as shown in Figure 1. The input layer parameters (C) are used as the distributed representations in our experiments. Whereas, the output layer parameters (U) are used as an atemporal compass to align these distributed representations. Hence, the output layer parameters are frozen and unchanged for further training, such that: $U = U_1^t = U_2^t = U_3^t = ...U_n^t$. The diachronic representations for each temporal slice in the corpus (D_i^t) are then obtained by fine-tuning the input layer parameters (C_i^t) on its corresponding temporal slice (D_i^t).

The fine-tuning for each temporal slice (D_i^t) is resumed with the representations from the previous temporal slice (C_{i-1}^t). This is done by initializing the input parameters (C_i^t) with the already fine-tuned input parameters from the previous temporal slice (C_{i-1}^t). Where the parameters for the newly acquired words in (D_i^t) are initialized randomly. This ensures that the diachronic model captures the lexical development in a cognitively plausible incremental way, following the paradigm of curriculum-learning in children. Formally, given a slice D^t the training procedure for an input $(w_k, \gamma(w_k))$ is defined as the following optimization problem:

$$max\Big(\log P(w_k|\gamma(w_k))\Big) = \sigma\Big(\vec{u}_k \cdot \vec{c}^{\,t}_{\gamma(w_k)}\Big) \quad (3)$$

carried out on C^t where the function σ is calculated using negative sampling, $\gamma(w_k) = (w_1, w_2, ...w_M)$ is the set of M words that appear in the context of w_k ($\frac{M}{2}$ being the size of the window), $\vec{u}_k \in U$ is the atemporal target embedding of the word w_k and $\vec{c}^{\,t}_{\gamma(w_k)}$ is the mean of the temporal context embeddings.

We train separate models for child-speech and child-directed speech, each providing distributed word representations independent of each other. We use these models to compare the lexical development in child-speech and child-directed speech. We also train both these models in both, the proposed incremental manner, and in a non-incremental way (Carlo et al., 2019) for ablation purposes. We train the models with three different random initialization seed values, and report the results averaged across the random seeds.

Data	Semantic	Syntactic
Child-speech	141	347
Child-directed speech	184	597
Combined	**126**	**335**

Table 1: The number of common probe-words across all the 19 months of data.

4 Analysis

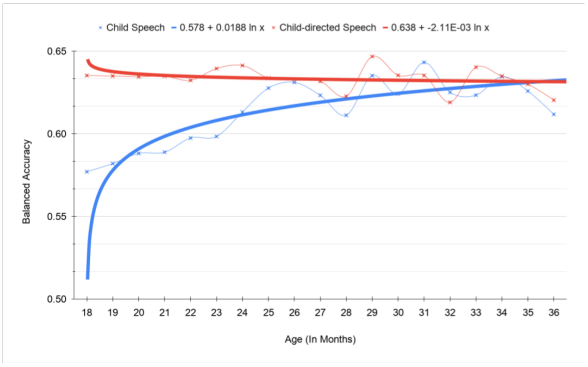
Given the parameterized representational nature of our model, it can be used to study the various empirical and qualitative aspects of lexical development in children. Here, we show the usability of our model by performing three experiments, and qualitative analysis[4] of its representations. Similar to the previous studies, we use a set of syntactic and semantic probe words (that occur frequently in child-speech) for this purpose. These probe words are derived from the MacArthur-Bates Communicative Development Inventory (MCDI). We obtain the probe words from the data used in Huebner and Willits (2018a).

We consolidate the vocabularies and obtain the common words appearing across all of the temporal slices. From this set of common words, we only consider the ones that are a part of the previously obtained semantic and syntactic probe words, where the rest of the words are discarded for analysis. The final set of syntactic probe words is classified into eight part-of-speech categories, and the final set of semantic probe words is classified into 24 abstract semantic categories. The statistics for the final set of probe words are given in Table 1.

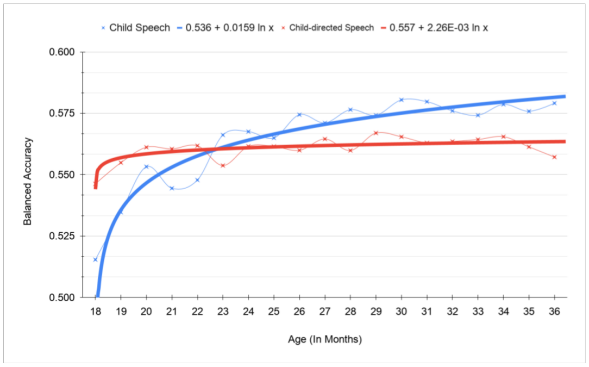
4.1 Lexical Category Learning

Performing lexical categorization is an important aspect of lexical development. The representational nature of our model allows it to perform categorization with any vector-similarity-based measure. We borrow the task of lexical categorization from Huebner and Willits (2018b). As the probe words are divided into well-defined syntactic and semantic categories, we quantify the category learning ability of our model using a balanced accuracy measure (Huebner and Willits, 2018b).

[4]We report the details for the qualitative analysis in Appendix A.

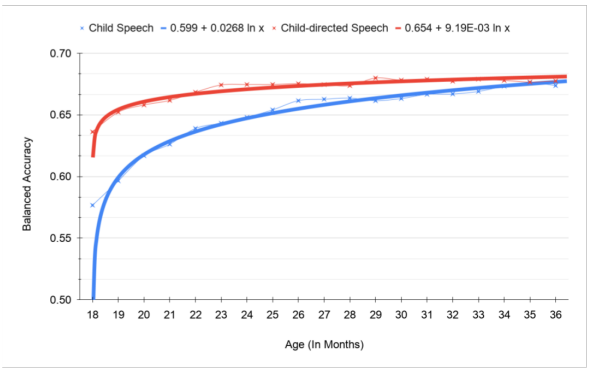


(a) Semantic Categorization

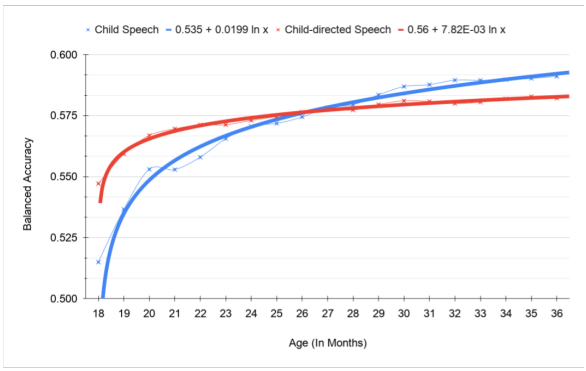


(b) Syntactic Categorization

Figure 2: Month-wise Balanced Accuracy scores for the lexical categorization experiment with the non-incremental diachronic representations for the child-speech and child-directed adult speech.



(a) Semantic Categorization



(b) Syntactic Categorization

Figure 3: Month-wise Balanced Accuracy scores for the lexical categorization experiment with the incremental diachronic representations for the child-speech and child-directed adult speech.

We use signal detection theory to calculate the measure of interest. For each probe word, a comparison is done with the other probe words in their own category as well as the ones in different categories. We use cosine-similarity measure $cosine(w, w')$ for this purpose. Two words are classified into the same category only if $cosine(w, w') \geq r$, where $r \in (0, 1)$ is a threshold value. Each correct classification is recorded as a hit and the incorrect classification is recorded as a $miss$. This is finally used to calculate a balanced accuracy measure (BA):

$$BA = \frac{TPR + TNR}{2} \quad (4)$$

where TPR (true-positive rate) is the $sensitivity$ and TNR (true-negative rate) is the $specificity$. For each temporal slice, the threshold value (r) is calculated to maximize the balanced accuracy in an iterative way (with a step-size of $1e-3$).

We perform the lexical categorization task with both non-incremental (Carlo et al., 2019) and incremental modeling (ours) approaches. We calculate the balanced accuracy measure for both the syntactic and semantic probe words for each month. This gives us a trajectory of lexical category learning in child-speech as shown in Figure 2 and Figure 3. We also plot the month-wise balanced accuracies for the child-directed adult speech, which gives us a trajectory of the lexical knowledge present in it. We model these trajectories by applying a temporal logarithmic fit[5] over the balanced accuracies values in the following manner:

$$BA = \alpha + \beta \times log_e(t) \quad (5)$$

where BA is the balanced accuracy, α is the intercept, and β is the log-curve coefficient.

Overall, we observe that the lexical categorization knowledge in children increases logarithmically over time, eventually saturating around the almost-constant level of existing categorization

[5]The final curve fit equations are given in the legends of Figure 2 and Figure 3.

Category	Word frequency in:	β_f	β_t	$\epsilon_{w_i}^{(t)}$
Syntactic Probe Words	Child-speech	-0.272	-0.047	1.008
	Child-directed speech	**-0.096**	-0.104	0.714
Semantic Probe Words	Child-speech	-0.322	-0.141	1.156
	Child-directed speech	**-0.224**	-0.182	1.058

Table 2: The results for the linear mixed random-effects models fitted on semantic change values $\Delta^{(t)}w_i$ in child-speech, with respect to the word frequencies in child-speech and child-directed speech.

knowledge in adults (child-directed speech). For the semantic categorization, we observe that the incremental model gives a maximum balanced accuracy[6] of 0.6738 (t=36 months), and the non-incremental model gives a maximum balanced accuracy of 0.6118 (t=36 months). Similarly, for the syntactic categorization the maximum balanced accuracy scores are 0.5911 (t=36 months) and 0.5791 (t=36 months) for the incremental and non-incremental models respectively.

Hence, the incremental model shows a significant improvement of 10.13% for the semantic categorization, and 2.07% for the syntactic categorization. The incremental model also shows smoother and monotonic balanced accuracy trajectories. This demonstrates the importance of the cognitively motivated curriculum learning method in our model, where the incremental model captures better lexical knowledge in its distributed representations than the non-incremental model.

4.2 Word Frequencies and Lexical Development

Word-frequencies have been extensively analyzed under various aspects of child-speech and child-directed speech in many previous works (Ambridge et al., 2015). Diachronic word embeddings have also been used to postulate laws mapping frequency-based measures to the historical semantic shifts of words. One such law by Hamilton et al. (2016) states that frequent words change more slowly. This can be formally expressed as:

$$\Delta w_i \propto f(w_i)^{\beta_f} \quad (6)$$

where Δw_i is the rate of semantic change, $f(w_i)$ is the frequency of the word w_i and β_f is a negative power as per the relation.

Given the word-frequency data, and the distributed word representations from a diachronic model, similar effects of word-frequencies can be inspected for lexical development in children. In order to demonstrate this, we borrow a modeling approach by Hamilton et al. (2016). Semantic change for each word at consecutive time steps $(t, t+1)$ can be calculated as:

$$\Delta^{(t)}w_i = 1 - cosine(w_i^{(t)}, w_i^{(t+1)}) \quad (7)$$

In the context of child speech, this value can be looked upon as the update in the mental representation of the word in the child's mental lexicon. The trajectory of a word's semantic change values can then be thought of as the process of acquiring (learning and grounding) the meaning of that word.

Following Hamilton et al. (2016)'s approach, we log transform and normalize the $\Delta^{(t)}w_i$ values. The $\Delta^{(t)}w_i$ values that are less than 0.05 are not considered in order to maintain numerical stability in the logarithm and to ignore the insignificant changes in the representations. These new values are denoted as $\bar{\Delta}^{(t)}w_i$. We then fit a linear mixed random-effects model in the following manner:

$$\bar{\Delta}^{(t)}w_i = \beta_f log\Big(f^{(t)}(w_i)\Big) + \beta_t + z_{w_i} + \epsilon_{w_i}^{(t)} \quad (8)$$

where β_f and β_t are fixed effects for $frequency$ and $time$ respectively, z_{w_i} is the random intercept and $\epsilon_{w_i}^{(t)}$ is the $error$ term.

We use the representations from the incremental model to obtain the $\Delta^{(t)}w_i$ values for child-speech. We fit separate linear mixed random-effects models for frequency data from child-speech and child-directed speech as shown in Table 2.[7] Similar to the findings of Hamilton et al. (2016), we find that β_f takes negative values across both syntactic and semantic probe words, and word frequencies from both the child-speech and the child-directed speech. While it can be argued that the $word2vec$ model's

[6] calculated with the representations from child-speech.

[7] all the obtained model fits are statistically significant with p-value < 0.05

dependence on word co-occurrence statistics might superficially induce negative β_f values for frequencies from child-speech, the negative β_f values for frequencies from the child-directed speech are independently obtained (given that the model trained on child-speech is not exposed to the data from child-directed speech at any point of time). Hence, we majorly focus on the β_f values from child-directed speech, which are obtained from the input word frequencies to the children. These negative β_f values are, in general, in good agreement with all the previous studies on the role of input word frequency in word acquisition (Ambridge et al., 2015). Where it is known that higher single-word frequencies are usually associated with quicker word acquisition (which translates to smaller semantic change values $\Delta^{(t)}w_i$ with respect to the temporal slices of the word exposure).

While the β_f values for semantic probe words are significantly negative as expected, the β_f values for syntactic probe words, although negative, are slightly close to 0. While this is only in a weak agreement with most of the previous studies, it is important to note that these studies inspect the role of input word frequencies with respect to specific syntactic constructs and categories (Ambridge et al., 2015). Whereas our results are representative of all the syntactic words in general.

4.3 Representational Similarity Analysis (RSA)

Lexical acquisition in children is pragmatized by the child-directed adult speech (Clark, 2017). While our results from the Lexical Categorization task (Section 4.1) implicitly depict the lexical knowledge transfer from adult to child, a more fine-grained analysis can be performed by directly comparing the distributed representations for child-directed speech and child-speech. Recent work in natural language processing research has focused on using Representational Similarity Analysis (RSA) (Laakso and Cottrell, 2000; Kriegeskorte et al., 2008) for various interpretability studies (Abnar et al., 2019; Gauthier and Levy, 2019; Lepori and McCoy, 2020; Merchant et al., 2020).

We use RSA to compare the diachronic representational geometries of child-speech and child-directed speech. Following the settings used by Lepori and McCoy (2020), we use Spearman's correlation (ρ) as the similarity metric (sim). For each time-step (i.e. month-wise), we first obtain the individual geometries for the corresponding representations for child-directed speech and child-speech by using the dissimilarity metric: $1 - sim$. For a fair comparison with the results from Section 4.1, we only use the similar set of semantic and syntactic probe words to obtain the representational geometries. The final similarity value between the representational geometries is then obtained by using the similarity metric (sim).[8]

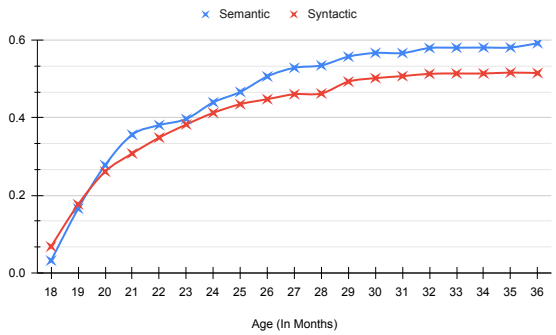


Figure 4: The similarity values between the representations for child-directed speech and child-speech (Spearman's correlation (ρ) vs. age in months).

The trajectory observed for representational similarities (Figure 4) across both the semantic and syntactic probe words matches with that of the balanced accuracy from the Lexical Categorization task (Figure 3). Hence, the incremental diachronic representations successfully capture the fine-grained dynamics of lexical acquisition, which ultimately translates to a higher-level of lexical processing.

5 Conclusion

In this paper, we explore the usability of diachronic distributed word representations towards modeling lexical development in children. While all the related previous works use distributed representations with child-directed speech only,[9] we also obtain the distributed representations of child-speech. This allows us to model the lexical development in children in a more direct way. Through an ablation experiment, we demonstrate the effectiveness of our cognitively motivated incremental learning diachronic model in capturing abstract lexical knowledge in noisy child-speech. We show the usability of our model across various dimensions of the study

[8] all the obtained correlation values are statistically significant with p-value < 0.05

[9] To the best of our knowledge.

of lexical development through multiple representative empirical and qualitative analyses.

Our experiment with the lexical categorization task reveals the trajectories of semantic and syntactic knowledge acquisition in children. Our experiment with the linear mixed-effect modeling of diachronic representational-changes displays the role of input word frequencies in word acquisition. Further, we also perform a fine-grained analysis of lexical knowledge transfer with Representational Similarity Analysis of diachronic representations from child-speech and child-directed adult speech. Our qualitative analyses reveal the phenomena of grounding, abstraction, categorization, and word associations in the mental lexicon of children in an elegant and simple manner (Appendix A).

6 Future Directions

The demonstrated effectiveness and ease of usage of diachronic distributed word representations opens up multiple future directions of research in modeling lexical development. While this work only deals with the usability of our model for word-level studies, the diachronic distributed representations from our model can also be used to study the psycholinguistic development at other granularities as well. Representations for higher granularities (partial-words, syllables, etc.) can be obtained by applying any vector-decomposition method over these word representations. Similarly, the representations for lower granularities (phrasal, clausal, sentence-level, etc.) can also be obtained by applying various vector-pooling techniques over the word representations. Lexical development is usually a multimodal process, where various perceptual modalities are involved. As the data collection efforts for child-speech advance, one can incorporate embeddings from other modalities (phonemic embeddings, visual embeddings, etc.) in modeling the lexical development.

While we use a fairly recent diachronic word embedding model in this work (Carlo et al., 2019), the lexical modeling efficiency can be increased in parallel with the advances in diachronic word embedding modeling. Further, handling the challenges like partial words, low vocabulary size, lesser training data, etc. can be a good research direction as well. In the future, we plan to extend this work to other languages, using data collected with subjects from a diverse demography.[10]

[10]The code and data used for this work will be made publicly available post publication.

References

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203, Florence, Italy. Association for Computational Linguistics.

Ben Ambridge, Evan Kidd, Caroline F Rowland, and Anna L Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of child language*, 42(2):239–273.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Edith Laura Bavin and Letitia R. Naigles. 2017. *The Cambridge handbook of child language*. Cambridge University Press.

Leonard Bloomfield. 1933. *Language*. Holt, Rinehard and Winston.

Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. 2019. Training temporal word embeddings with a compass.

Eve V Clark. 2017. Lexical acquisition and the structure of the mental lexicon. In *Oxford Research Encyclopedia of Linguistics*.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Abdellah Fourtassi. 2020. Word co-occurrence in child-directed speech predicts children's free word associations. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 49–53, Online. Association for Computational Linguistics.

Abdellah Fourtassi, Isaac Scheinfeld, and Michael Frank. 2019. The development of abstract concepts in children's early lexical networks. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 129–133, Minneapolis, Minnesota. Association for Computational Linguistics.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 529–539, Hong Kong, China. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Philip Huebner and Jon Willits. 2018a. Order matters: Developmentally plausible acquisition of lexical categories.

Philip A. Huebner and Jon A. Willits. 2018b. Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*, 9.

Hang Jiang, Michael C Frank, Vivek Kulkarni, and Abdellah Fourtassi. 2020. Exploring patterns of stability and change in caregivers' word usage across early childhood.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2014. Statistically significant detection of linguistic change.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Aarre Laakso and Garrison Cottrell. 2000. Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology*, 13(1):47–76.

Jill Lany and Jenny R. Saffran. 2010. From statistics to meaning: Infants' acquisition of lexical categories. *Psychological Science*, 21(2):284–291. PMID: 20424058.

Michael Lepori and R. Thomas McCoy. 2020. Picking BERT's brain: Probing for linguistic dependencies in contextualized embeddings using representational similarity analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3637–3651, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. Third Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Efficient estimation of word representations in vector space.

David E. Rumelhart and James L. McClelland. 1986. *Parallel distributed processing: Explorations in the microstructure of cognition, vol.1: Foundations*. The MIT Press.

Kristen Syrett and Jeffrey Lidz. 2010. 30-month-olds use the distribution and meaning of adverbs to interpret novel adjectives. *Language Learning and Development*, 6(4):258–282.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188.

Laurens van der Maaten and Geoffrey Hinton. 2008. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Shirui Wang, Wenan Zhou, and Chao Jiang. 2020. A survey of word embeddings based on deep learning. *Computing*, 102(3):717–740.

Yating Zhang, Adam Jatowt, Sourav Bhowmick, and Katsumi Tanaka. 2015. Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 645–655, Beijing, China. Association for Computational Linguistics.

A Qualitative Analysis

A.1 t-SNE Visualization

A significant advantage of using distributed representational models is their suitability for qualitative analyses. Given a high-dimensional vector representation space, one can apply various dimensionality reduction algorithms to map it to a two-dimensional space with minimum errors. Which in turn allows one to visualize the representations on a 2D plot.

We use a t-SNE dimensionality reduction algorithm (van der Maaten and Hinton, 2008) on the incremental representations from the child-speech data at Age = 18 months and Age = 36 months. We use a perplexity value of 19 for the t-SNE algorithm. We use the mean vector in each probe-word category to get its $centroid$. To exclude any extreme outliers from the plots, we use Chebyshev's inequality and limit the X-coordinates with a value of $k = 8$ standard deviations. We observe several qualitative drifts that emerge in the representations obtained at Age = 36 months as compared to those obtained at Age = 18 months (Figure 5).

For the semantic categories (Figure 5d), the words belonging to the categories containing living creatures: $plant$, $mammal$, $insect$, and $bird$ cluster together. The related categories of $household$ and $bathroom$, and the food-related categories: $fruit$, $dessert$ and $drink$ appear together as well. Words in the $clothing$, $body$, and $furniture$ occupy a distinct portion of the space, hinting towards the emergence of grounded word meanings (of clothing) to locations (near furniture like closet and mirror) and usage (over the body). Clusters day and $times$ drift closer as their constituent words are frequently used together. All the food categories appear adjacent to the $kitchen$ category in the space.

Similar strikingly visible patterns are observed with syntactic categories as well. Unlike the syntactic representations at Age = 18 months (Figure 5a), the syntactic representations at Age = 36 months show well-clustered categories (Figure 5b). The two major categories: $noun$ and $verb$ become almost linearly separable. Their related categories are sorted accordingly as well. The $pronouns$ and $adjectives$ are placed in the upper half of the plot, occupied by the $noun$ category. On the other hand, $adverbs$ appear in the bottom half of the plot, which is occupied by the $verb$ category. The remaining neutral categories: $determiners$, $interjections$ and $adpositions$ are well placed at the boundary of the $noun$ and $verb$ clusters.

A.2 Nearest Neighbors

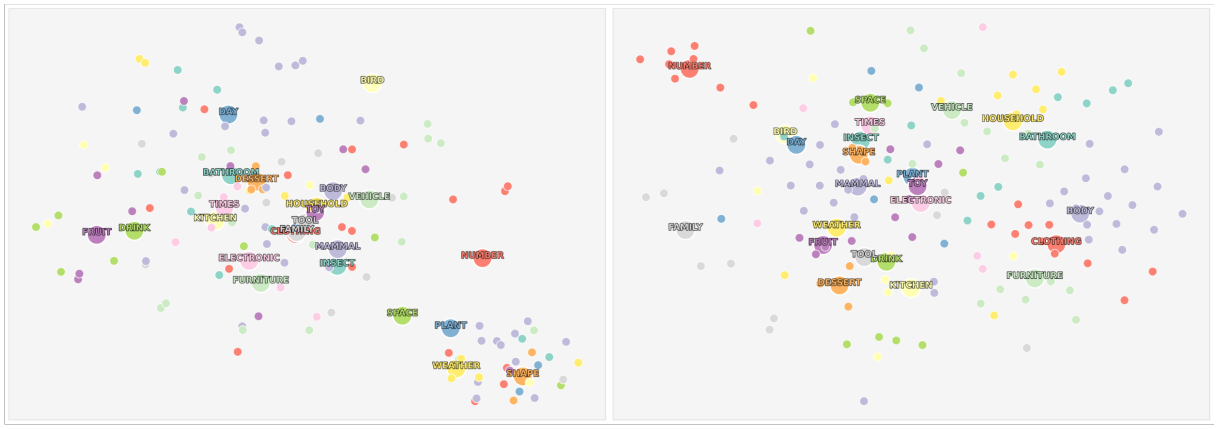
Another fine-grained approach towards the qualitative analysis of distributed representations is that of observing the nearest neighbors of particular data points.

We note the k-nearest neighbors[11] ($k = 3$) for a target word from each semantic (Table 3) and syntactic category (Table 4). While the neighbors at Age = 18 months are a bit random, the neighbors at Age = 36 months appear to be more systematically relevant, either by belonging to the same category (example: $zoo \rightarrow \{store, school\}$), or by showing certain abstract free word-associations (example: $tea \rightarrow \{cup, milk\}$; $wet \rightarrow \{dirty, diaper\}$).

[11] We use cosine-distance to find the nearest neighbors.

(a) Syntactic (Age = 18 months)

(b) Syntactic (Age = 36 months)



(c) Semantic (Age = 18 months)

(d) Semantic (Age = 36 months)

Figure 5: t-SNE visualizations of the probe word categories at Age = 18 months and Age = 36 months.

Word	Neighbours at 18 months	Neighbours at 36 months	Category
towel	diaper, floor, neck	diaper, paper, blanket	BATHROOM
duck	cake, hi, bird	square, bird, boat	BIRD
tummy	got, your, hurt	finger, tongue, head	BODY
tie	touch, break, try	wear, pull, push	CLOTHING
today	maybe, move, many	camera, kids, bus	DAY
cookie	cookies, still, happy	cookies, breakfast, strawberry	DESSERT
tea	am, heavy, ready	coffee, cup, milk	DRINK
telephone	talk, doctor, blow	phone, couch, plate	ELECTRONIC
mom	talking, dinner, sleeping	dad, mommy, six	FAMILY
strawberry	apple, yep, cheese	apple, banana, cheese	FRUIT
table	under, chair, sitting	floor, couch, wall	FURNITURE
window	fell, running, said	door, kitchen, spider	HOUSEHOLD
spider	fire, wall, moon	window, sun, bear	INSECT
spoon	floor, hey, side	fork, bowl, cup	KITCHEN
tiger	eight, blue, green	dinosaur, chicken, lion	MAMMAL
two	three, four, can	many, four, five	NUMBER
tree	climb, nap, stand	wall, climb, square	PLANT
square	ooh, ah, funny	circle, butterfly, big	SHAPE
sun	pencil, door, wash	moon, snow, dog	SPACE
night	said, warm, hey	morning, time, day	TIMES
vacuum	end, running, am	bike, careful, room	TOOL
toy	pants, running, sweater	game, block, lion	TOY
truck	fire, man, drive	tractor, plane, car	VEHICLE
snow	heavy, plane, egg	sun, wow, grass	WEATHER

Table 3: Nearest semantic neighbours for $k = 3$ at the first and last temporal slice.

Word	Neighbours at 18 months	Neighbours at 36 months	Category
wet	water, diaper, baby	dirty, diaper, hurt	ADJ
with	different, game, morning	game, lap, help	ADP
where	she, are, how	yes, who, how	ADV
your	hands, tummy, feet	you, yours, okay	DET
yes	bear, give, blanket	what, where, yep	INTJ
zoo	bus, mommy, five	store, school, party	NOUN
yours	who, touch, noise	pen, coffee, candy	PRON
write	set, ready, touch	draw, pencil, pen	VERB

Table 4: Nearest syntactic neighbours for $k = 3$ at the first and last temporal slice.