Modelling Children's Sentence Recall using an Encoder-Decoder Network

Modelling Children’s Sentence Recall using an Encoder-Decoder Network

Daniel Freudenthal (D.Freudenthal@liverpool.ac.uk)
Julian M. Pine (JPine@liverpool.ac.uk)
School of Psychology, University of Liverpool

Colin Bannard (colin.bannard@manchester.ac.uk)
Department of Linguistics and English Language, University of Manchester

Abstract

Elicited imitation is a widely used method for testing a child's knowledge of a language for scientific or clinical purposes. A child hears an utterance and is asked to repeat what they have heard. While it is assumed that their fluency or speed in doing so is contingent on their linguistic competence, little is known about the cognitive mechanisms and/or representations involved. To explore this, we train an encoder-decoder model, consisting of recurrent neural networks, to encode and reproduce a corpus of child-directed speech and then test its performance on the experimental task of Bannard and Matthews (2008). In that study pre-school children were asked to repeat high- and low-frequency four-word sequences in which the first three words were identical (e.g., *sit in your chair* and *sit in your truck*) and the final words and bigrams were closely matched for frequency. We find that like those children our model makes more errors on the initial three words when they are part of a low-frequency than a high-frequency sequence, despite the fact that the words being repeated are identical. We explore why this might be and pinpoint some possible similarities between the model and child language processing.

Keywords: sentence repetition; language development; deep learning; phrase frequency.

Introduction

Elicited imitation, also referred to as sentence repetition or sentence recall, is a widely used method for testing a child's knowledge of a language. A child hears an utterance being produced, typically in a recording, and is asked to repeat what they have heard. The child’s fluency and/or speed in doing so is taken as an indicator of their knowledge of linguistic (typically lexical or grammatical) properties of the sentence.

As well as being used to address a range of theoretical questions regarding the acquisition of language, over the last two decades elicited imitation has found application in clinical diagnosis. Conti-Ramsden et al (2001) explored the utility of sentence repetition in identifying children with Developmental Language Disorder. They found it to be a more useful measure than three other more widely-used tasks. Since then, the method has been deployed across multiple languages (see Rujas et al, 2021 for a recent survey).

It is widely assumed that the child's ability to remember and then reproduce a sentence is affected by their linguistic ability. According to Potter (2012; p.5) "immediate recall of a sentence (like longer-term recall) is based on a conceptual or propositional representation of the sentence…having understood the conceptual proposition in a sentence, one can simply express that idea in words, as one might express a new thought." However, there is limited understanding of the mechanisms involved or the encodings used. One way that this gap can be addressed is of course via human experiments - looking at how well children are able to represent and recall different kinds of sentences. Another, complementary, approach is to create a computational model of the process. In this work we take a step in this direction.

Our starting point for a model of elicited imitation is the encoder-decoder network. Such networks are used in a range of natural language processing tasks to generate utterances conditional on an input. First the encoder builds a representation of an input utterance, including the order of the words. This (whole sentence) representation is then passed on to the decoder, whose task it is to generate a sentence conditioned on the encoded input. For example, in machine translation an utterance is generated in a target language conditional on an utterance in a source language. In dialogue systems an utterance is generated conditional on the previous turn in the dialogue.

Our goal in this paper is to model the process of sentence repetition. The input to our network is the sentence heard by the child. The job of the network is to encode and then reproduce this sentence. We are interested in whether the kinds of representations that encoder-decoder networks develop in response to this task can provide a useful model of those used by children. Our window onto human representations is the errors that speakers make. However, to a fully trained network with unlimited capacity, direct imitation is a trivial task, much as the sentence repetition task is trivial for adult speakers. We therefore impose constraints on the representational capacity of the model.

Our question is whether the circumstances under which our model makes errors are the same as those under which children make errors. We seek to replicate an effect reported by Bannard and Matthews (2008). In this study 2- and 3-year-old children repeated pairs of four-word sequences that were identical except for the final word, where one sequence was high frequency (*a drink of milk*) and one was low frequency (*a drink of tea*) in child-directed speech. The frequency of the final words and bigrams were also matched. Children were found to be more fluent and quicker in producing the identical first three words when those words were part of a high-frequency sequence. This finding, where the error rate for a single sequence of words varies depending on the context, provides an interesting test case for any model of production.

While encoding and decoding can be done by a range of different temporal-learning networks, the model type we deploy is a kind of recurrent neural network - the long short term memory model (LSTM; Hochreiter & Schmidhuber, 1997). Unlike in simple recurrent networks which have been widely deployed for cognitive modelling purposes, in LSTMs the flow of information across time is controlled by memory gates, the behaviour of which is also learned from data. Recurrent neural networks with gating are the basis of a popular cognitive model of working memory (O'Reilly & Frank, 2006) with recent work having identified neural signatures of gating (Rac-Lubashevsky & Frank, 2021).

Methods

Model Architecture

Encoder-decoder models, also known as Seq2Seq models, consist of two components: an encoder and a decoder. The precise function of these components depends on the task being performed. In machine translation, the model is trained on pairs of translationally equivalent sentences. The task of the encoder is to build a representation of the source utterance, including the order of the words it contains. This (whole sentence) representation is then passed on to the decoder, whose task it is to produce a translation of the source utterance in the target language. The decoder is presented with the encoder's representation of the source utterance (as well as a start-of-utterance marker) and starts to produce words in the target language. Importantly, the words that are produced by the decoder are fed back to it. The decoder thus blends the encoder's representation of the source utterance with the representation of the words it has produced so far[1] to shape the words it produces downstream (see Fig. 1). This process continues until the decoder produces an end-of-utterance marker, or a maximum word limit is produced. Since learning is supervised, the decoder's production can be compared to the target utterance, to generate an error signal that is used to adjust weights throughout the model.

We implement our encoder and decoder using an LSTM network. Both the encoder and decoder can be equipped with so-called embedding layers, which can provide a dense representation capturing the semantic (dis)similarity between words. The use of embedding layers representing distributional semantics represents an improvement over traditional language modelling methods (such as n-gram models) as it enables the computation of linguistic representations over classes of distributionally similar items.

For the current simulations we used a Seq2Seq model in which both the encoder and decoder contained a single layer LSTM network. We investigated how LSTM capacity affects model performance by running models with differing numbers (30, 40, 50, 60) of hidden cells or dimensions. Both the encoder and decoder were equipped with an embedding layer of 50 dimensions. Embeddings were learned during training and thus specific to the model's input which consisted of child-directed speech. Code for all experiments is at github.com/cbannard/sentrep_cogsci22.

An important difference between the standard use of the Seq2Seq model and our simulations is that, unlike in machine translation, where the source and target utterance are translational equivalents in different languages (which may be of different length), our source and target utterances are pairs of identical English utterances. This may seem a trivial task from a machine translation perspective. However, it is worth bearing in mind that the encoder and decoder are independent, and the encoder needs to learn to represent the input it receives in a way that is sufficiently fine-grained for the decoder to infer both the identity and the order of the words that the encoder received. The encoder and decoder roughly map onto the distinct processes of comprehension and production, and comparing the model's and children's performance on sentence repetition may provide us with insights into the representations involved in these processes.

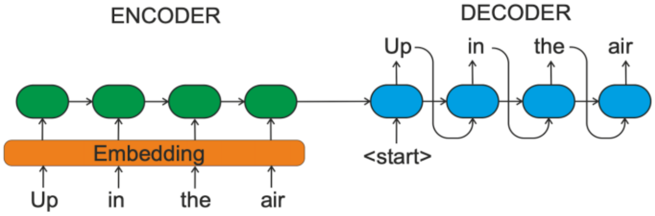

Figure 1: An Encoder-Decoder (Seq2Seq) network

Model Input

Our model was trained on English child-directed speech obtained from CHILDES. Input consisted of a mix of UK and US English corpora. We performed minor filtering on the CHILDES files (we removed punctuation and markup). For information on the corpora and preprocessing employed see the github repository. The total amount of input approximated 2 million utterances, restricted to a maximum length of 15 words. We reduced the vocabulary to words that occurred a minimum of 5 times. Out-of-dictionary words in the input were replaced with a random token. Models were trained for a total of 500 epochs of 10,000 utterances each. The input for each individual epoch was sampled randomly from the input corpus and split in an 80:20 ratio for training and validation. Models were tested (and responses recorded)

[1] Though learning can be aided by feeding back the words from the target utterance, a process known as Teacher Forcing.

on the experimental stimuli (see below for details) every 10 epochs.

Experimental Stimuli

The stimuli were 13 pairs of high-frequency (e.g., *go to the shop*) and low-frequency sequences (e.g., *go to the top*) listed in Bannard and Matthews (2008). The high- and low-frequency items from each pair had the same first three words and final words and bigrams were of equivalent frequency.

Results

Model performance (error rates on the first three words) on the stimuli is shown in Fig. 2. Models were run with 30, 40, 50 or 60 hidden dimensions in both the encoder and decoder LSTMs, and averaged over 10 model runs each. As can be seen in Fig 2, all models learn to repeat the stimuli with high accuracy. Models with more hidden dimensions do so more quickly, and reach better overall accuracy levels than models with fewer hidden dimensions. It is clear from Fig. 2 that all models go through a stage where they show better performance on the first 3 words of the high-frequency than the low-frequency sequences. The degree of separation, as well as the length of the period during which there is separation, is larger for models with lower capacity. Overall, the degree of separation is in a similar range to that reported by Bannard and Matthews (~10% for the two-year-olds, and ~5% for the 3-year-olds), though not necessarily at the same overall error rates (68% on the low frequency items for the two-year-olds, and 35% for the three-year-olds). Nevertheless, the fact that the models show separation between the two sets of stimuli across a range of 100-200 epochs suggests that the model architecture captures some aspects of the difference between the two sets of stimuli that children are sensitive to.

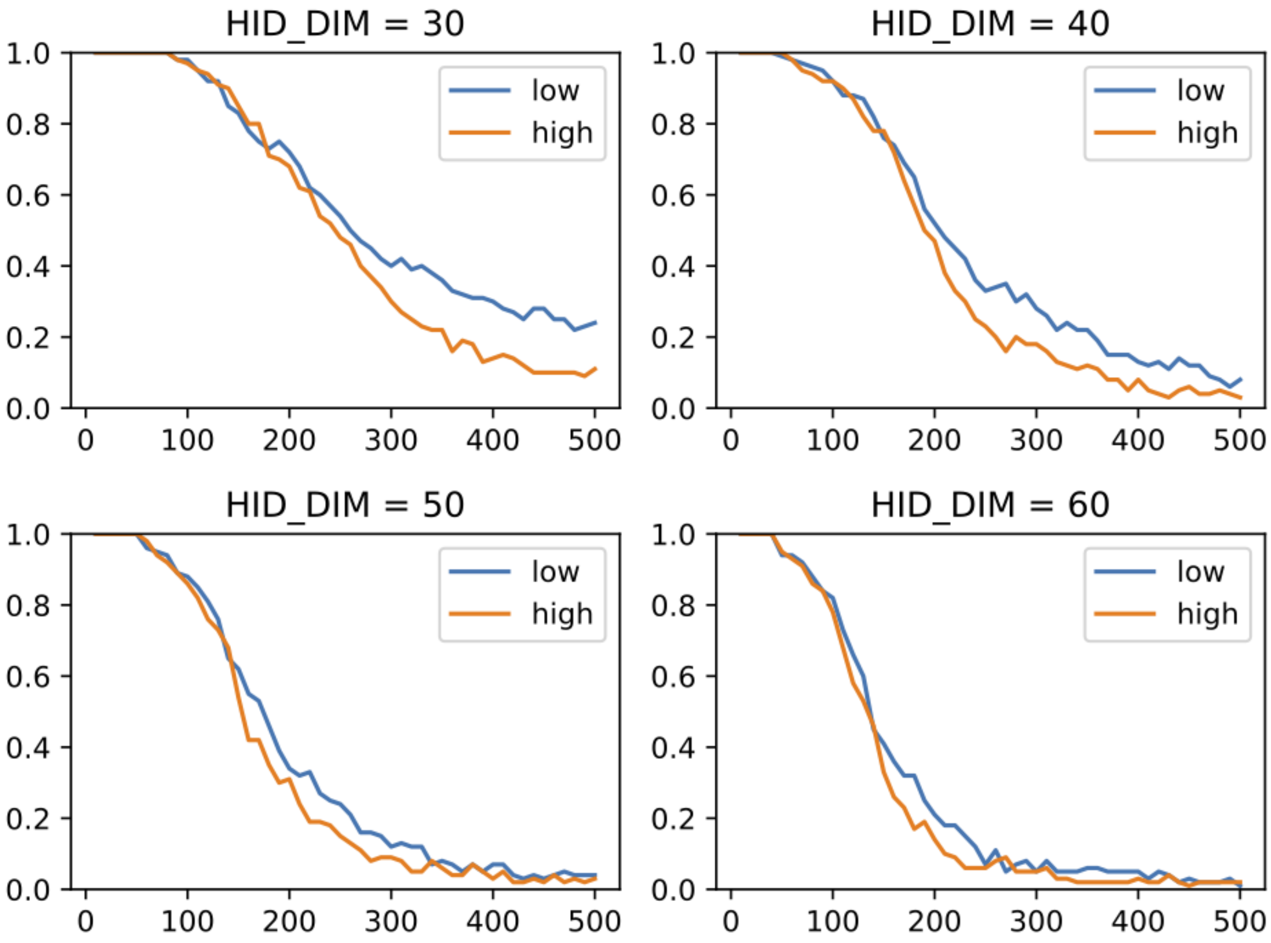

Figure 2: Model performance (3-word error rate) on the Bannard and Matthews (2008) stimuli with high and low frequency.

Error Analysis

We next look at the position and types of error that the model makes. For this analysis we pooled the data from all (10) models with hidden dimensions of 40 and 50 at epochs 150, 200, 250 and 300. This sample reflects a range of intermediate training stages and capacity levels. Since there are 26 stimuli (13 pairs) in the Bannard and Matthews set, this makes up a set of 2080 productions. The three-word error rate in the combined set is 40% for the low-frequency sequences, and 34% for the high-frequency sequences.

The results shown in figure 2 include only errors on the first three words - the parts of the sequences that are identical in the high- and low-frequency four-word sequences. In the high-frequency sequences, these earlier words must somehow be protected from error by the larger sequence and the downstream target word. Take, for example, the sequence pair *when we go out* and *when we go in*. For an example model the first, high-frequency sequence is correctly produced, but for the second low-frequency sequence we see the sequence *we we go in.* It appears that the higher probability starting word *we* has intruded and is repeated. However, somehow in the word sequence *when we go out* the higher frequency of the target protects the production from this intrusion as early as the first word. In fact, errors seem to be made with highest frequency in earlier positions of the 2080 productions, 451 have errors in first position, 428 have errors in second position and 229 have errors in third position**.**

Errors in the productions can be loosely divided into two kinds - those that involve the production of a word that isn't in the target sequence and the production of a word from the target but in the wrong position. Errors of the second type can be separated into intrusions (the model produces a word that occurs at a later position in the target) and perseverations (the model produces a word that occurs at an earlier position in the target). Error rates for the different positions are shown in Table 1. For completeness we include errors on all four words.

Table 1: Error types by position.

	Target word in wrong place		Non-target word
	Intrusion	Perseveration	
1	288	0	163
2	183	74	171
3	36	92	101
4	0	172	405

Unsurprisingly, intrusion errors are most common in positions 1 and 2, while perseveration errors are more common in later positions. However, it is worth noting that intrusion errors make up a much larger proportion of errors in position 1 (64%), than perseveration errors do in position 4 (30%). Meanwhile non-target intrusions are most likely in this final position. We propose the following explanation for this pattern. All of these errors arise because highly activated words appear in place of less activated words. Words that are less activated early in the sequence remain so throughout the production and thus are omitted rather than delayed. Lexical repetition of otherwise preferred words is unlikely as repetition is rare in the training corpus. These two pressures result in target words occurring earlier than they should more often than they occur late.

One factor that could plausibly be implicated in this process, as well as in the appearance of non-target words is the relative frequency of different words and word combinations in the training corpus - it could be that errors occur in the utterances at exactly the points where the target word is least predictable or where another word (from a later part of the target or from outside the target) is most probable in that context. This latter kind of error is known elsewhere in cognitive psychology as a habit slip and has recently been reported in linguistic behaviour by adults and children (Bannard et al, 2019; McCauley et al, 2021).

The existence of habit slips in our model output was investigated by examining the relative frequency of the target and the produced sequence for each error. The analysis was carried out in the following manner. For every error, we considered a left and right context, with the left context defined as all positions from (and including) the start marker up to (and including) the position containing the error, while the right context consisted of all positions up to (and including) the end marker. For both contexts, we aimed to determine the maximum context length for which the resulting production (the error) was more probable than the target. Thus, for an error in position 1, we considered as the left context the unigram (word) probability of the produced word compared to the target, as well as the bigram probability of the relevant word in combination with the start marker (i.e., in utterance-initial position). If the bigram probability of the error was larger than the target, we assigned a context score (length) of 2. If the bigram probability of the target was larger, and the unigram (word) probability of the error was larger than the target, we assigned a context score of 1. If both the bigram and unigram probability of the target exceeded those of the error, we assigned a context score of zero, indicating that the error was not driven by the n-gram statistics of the input. The same procedure was followed for the right context and across positions. The rationale behind this procedure is that each error is described by two numbers that express how well the error is supported by the left and right n-gram statistics, with higher numbers reflecting larger supportive contexts. The upper limit of these numbers depends on position and type of context and ranges from 2 (pos1, left; pos4, right) to 5 (pos1, right; pos4, left). The lower limit is always zero. Fig. 3 shows the results of this analysis as a stacked bar chart.

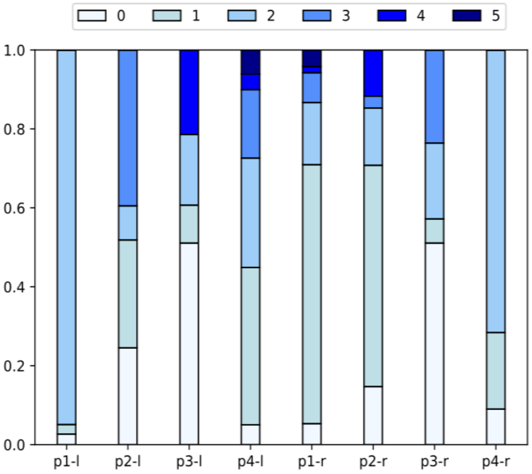

Fig 3: Length of supporting contexts for habit slips in different positions, e.g., P1-l = position 1, left context.

There are a number of things that stand out in Fig. 3. At the left context for position 1 (the leftmost bar), the overwhelming majority of errors is supported by a context of length 2, which represents the bigram probability in the context of the start marker (the maximum possible value for this context). This means that virtually all the errors in position 1, consist of words that are more frequent in utterance-initial position than the target word. By contrast, the largest stack for the right context of position 1 (bar #5), shows a majority maximum length of 1, meaning that the unigram (word) frequency of the error exceeds the target, but the error is not supported by larger n-grams. This pattern is consistent with the large number of intrusions in position 1. Many of the position 1 errors involve the intrusion (and repetition) of high- over low-frequency prepositions (e.g., ‘in in the air’ instead of ‘up in the air’). However, there are also a small number of position 1 errors that are supported by longer contexts (e.g., ‘put in your mouth’ instead of ‘sit in your truck’).

The natural comparison for position 1 is the errors seen in position 4. Nearly 60% of errors in position 4 are supported by a left context of length 2 or more, meaning that these errors result in relatively large n-grams that are more frequent in the input than the target n-gram. This is in marked contrast to position 1 where only around 30% of errors was supported by a right context of length 2 or more. Thus, position 4 errors are more influenced by left (preceding context) than position 1 errors are influenced by the right context (even though both have a maximum length of 5). Relatedly, the right context of position 4 (end marker) exerts less influence than the left context of position 1 (start marker). Taken together these analyses suggest that, as we might expect, as the utterance progresses the words that are produced become increasingly influenced by the words already produced and perhaps less by the encoded input.

Decoder analysis

The final stage in production for each word by our decoder involves the generation of a list of candidate words along with associated probabilities. The results above are all based on outputting the most frequent word at each step. However, further insight into the production processes of the model can be provided by looking at the full distribution. In Tables 2 and 3 we provide the relative production probability (Softmax), of the top 10 candidate words for the low-frequency target ‘up in the bath’ and its paired high-frequency sequence 'up in the air' (from a model with 50 hidden dimensions at 200 epochs). Table 2 shows an example of an intrusion in position 1: the word ‘in’ is considerably more common in utterance-initial position than ‘up’. In the continuation the model has inserted the word ‘water’ for the (semantically-related) ‘bath’. Both ‘in the water’ and ‘in the bath’ are relatively frequent in the input (at counts of 897 and 501 respectively). While the higher frequency of 'in the water' appears to give it the upper hand it is worth noting that ‘water’ is only the 3rd most frequent word in the ‘in the X’ frame, with the top two candidates being ‘box’ (1356) and ‘car’ (1179). It seems then that the existence of a semantically-overlapping phrase 'in the water' that is more frequent is what leads to this error.

Table 2: Relative production probabilities for low-frequency target ‘up in the bath’

Pos1	Pos2	Pos3	Pos4
in .59	in .68	the .99	water .65
up .21	up .09	in .0	bath .08
out .12	out .03	water .0	ball .02
you'll .01	of .02	some .0	money .02
there's .01	them .02	your .0	juice .01
more .01	water .01	of .0	bowl .01
on .0	the .01	front .0	door .01
here .0	into .01	more .0	paper .01
back .0	on .01	them .0	sun .01
two .0	from .01	paper .0	milk .0

By contrast it can be seen in table 3 that when producing the high-frequency sequence 'up in the air' the decoding process is protected from error from the start of the utterance on. Because the encoded target sequence and all of its subphrases are high frequency there is less opportunity for input-driven intrusions from high-frequency competitor words or substrings. The only apparent semantically-related word that has moderate production probability is the word 'sky' in the final word position, presumably supported by the medium frequency sequence 'up in the sky'.

Table 3: Relative production probabilities for the high-frequency target ‘up in the air’

Pos1	Pos2	Pos3	Pos4
up 0.58	in 0.61	the 0.98	air 0.06
in 0.3	up 0.22	in 0.0	trees 0.05
out 0.04	from 0.03	up 0.0	basket 0.03
two 0.01	into 0.02	of 0.0	sky 0.03
here 0.01	of 0.02	front 0.0	moon 0.02
you'll 0.01	out 0.01	from 0.0	stairs 0.02
there's 0.01	through 0.01	washing 0.0	kitchen 0.02
back 0.0	on 0.01	into 0.0	tree 0.02
from 0.0	with 0.01	through 0.0	road 0.02
down 0.0	the 0.0	your 0.0	pieces 0.01

Discussion

The goal of this work was to see whether an encoder-decoder network would show the same behaviour as children in an elicited imitation task (Bannard and Matthews, 2008).

In the experimental data, children were found to make more errors in repeating the first three words of low-frequency four-word sequences than they did when repeating exactly the same three words in high-frequency four-word sequences (i.e., where only the last word varied). We found that an LSTM-based encoder-decoder model showed exactly this behaviour. This was observed to be particularly the case when the model was given a limited number of hidden nodes and/or when it was trained for a limited number of iterations.

That an LSTM, or indeed any language model, should be better at repeating a more probable word sequence than a less probable one is to be expected. However, that it should be better at producing a sequence of three words when those words are part of a frequent four-word sequence than they are at repeating exactly the same words when they are part of a less frequent four-word sequence requires explanation.

In a simple LSTM language model (with no conditioning on an input utterance) the production of each word is conditioned only on those of the words produced so far that the network's gating mechanism deems relevant to downstream prediction (based on its training). In an encoder-decoder network, by contrast, each word is conditioned on this same information plus the representation of the input sentence which the decoder receives from the encoder and which is carried over at each step by the recurrent connections in the network. When producing the first three words of the sequences in our task, a simple LSTM would perform identically for high- and low-frequency four-word targets. Any difference between the two conditions must be due to the presence of the encoded sequence in the representation passed along via the recurrent network.

A likely explanation for the pattern that we see, then, is that the encoder finds it easy to encode high-frequency phrases - via an embedding space of unchanging size - even with limited representational resources (relatively few nodes in the hidden layer). The representation of the low frequency phrases, however, cannot be done so efficiently and the resulting representation may end up being dominated by component words or substrings. A decoding process conditioned on the former (high-frequency) encoding is likely to correctly output the target, while one conditioned on the encoding of the low-frequency string may make lexical selection errors. These errors are no longer produced once more representational resources (additional hidden nodes) or more training are provided.

The process that leads to the errors in our model provides a potentially useful approximation to what happens in child sentence repetition. Working memory is assumed to involve the allocation of resource-limited attention to long-term memory representations (D'Esposito & Postle, 2015). This process is in some ways analogous to the encoding process - a hidden layer linked to an embedding space - as seen in our model. We have only modelled one experiment in this initial work - we anticipate that applying the model to a larger range of stimuli will offer further insight into the processes and representations involved in children's elicited imitation.

Acknowledgements

Daniel Freudenthal, Julian Pine, and Colin Bannard are members of the International Centre for Language and Communicative Development (LuCiD) at the Universities of Liverpool and Manchester, for which the support of the Economic and Social Research Council [ES/L008955/1 and ES/S007113/1] is gratefully acknowledged.

References

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological science*, *19*(3), 241-248.

Bannard, C., Leriche, M., Bandmann, O., Brown, C. H., Ferracane, E., Sánchez-Ferro, A., ... & Stafford, T. (2019). Reduced habit-driven errors in Parkinson’s Disease. *Scientific reports*, 9(1), 1-8.

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of child psychology and psychiatry*, *42*(6), 741-748.

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual review of psychology*, *66*, 115-142.

Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*. 9 (8): 1735–1780.

McCauley, S. M., Bannard, C., Theakston, A., Davis, M., Cameron-Faulkner, T., & Ambridge, B. (2021). Multiword units lead to errors of commission in children's spontaneous production:“What corpus data can tell us?*”. *Developmental science*. 24(6). e13125.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural computation, 18(2), 283-328.

Potter, M. C. (2012). Conceptual short term memory in perception and thought. *Frontiers in Psychology*, *3*, 113.

Rac-Lubashevsky, R., & Frank, M. J. (2021). Analogous computations in working memory input, output and motor gating: Electrophysiological and computational modeling evidence. PLOS Computational Biology, 17(6), e1008971.

Rujas, I., Mariscal, S., Murillo, E., & Lázaro, M. (2021). Sentence repetition tasks to detect and prevent language difficulties: A scoping review. *Children*, *8*(578).