

Learning to generalise but not segment an artificial language at 17 months predicts children's  
language skills 3 years later

Padraic Monaghan<sup>1</sup>

Seamus Donnelly<sup>2</sup>

Katie Alcock<sup>1</sup>

Amy Bidgood<sup>3</sup>

Kate Cain<sup>1</sup>

Samantha Durrant<sup>4</sup>

Rebecca L.A. Frost<sup>5</sup>

Lana S. Jago<sup>1,3</sup>

Michelle S. Peter<sup>6</sup>

Julian M. Pine<sup>7</sup>

Heather Turnbull<sup>7</sup>

Caroline F. Rowland<sup>8,7,9</sup>

<sup>1</sup>Lancaster University

<sup>2</sup>Australian National University

<sup>3</sup>Liverpool John Moores University

<sup>4</sup>Manchester University

<sup>5</sup>Edge Hill University

<sup>6</sup>Great Ormond Street Hospital

<sup>7</sup>Liverpool University

<sup>8</sup>Max Planck Institute for Psycholinguistics

<sup>9</sup>Radboud University

Corresponding author: Padraic Monaghan, Department of Psychology, Lancaster University,

Lancaster LA1 4YF; Email: [p.monaghan@lancaster.ac.uk](mailto:p.monaghan@lancaster.ac.uk)

### Abstract

We investigated whether learning an artificial language at 17 months was predictive of children's natural language vocabulary and grammar skills at 54 months. Children at 17 months listened to an artificial language containing non-adjacent dependencies, and were then tested on their learning to segment and to generalise the structure of the language. At 54 months, children were then tested on a range of standardised natural language tasks that assessed receptive and expressive vocabulary and grammar. A structural equation model demonstrated that learning the artificial language generalisation at 17 months predicted language abilities – a composite of vocabulary and grammar skills – at 54 months, whereas artificial language segmentation at 17 months did not predict language abilities at this age. Artificial language learning tasks – especially those that probe grammar learning – provide a valuable tool for uncovering the mechanisms driving children's early language development.

Keywords: vocabulary, grammar, artificial language learning, non-adjacent dependencies, generalisation, segmentation

## Introduction

In order to acquire language, children need to be able to build a vocabulary – that is, they must determine what are the meaningful units in speech. They must also develop an understanding of the grammar of their language, which requires finding the relations between those words in utterances (Bates et al., 1988; Brinchmann et al., 2019; Dale et al., 2000; Dixon & Marchman, 2007; Labrell et al., 2014). How children resolve these two learning tasks can be characterised as a “chicken-and-egg” problem (Childers et al., 2012; Gleitman, 1990; Gleitman et al., 2005): in order to acquire the meaning of a word, its grammatical role in a sentence has to be understood, yet, in order to understand the role of the word in the sentence, the grammar of the sentence needs to be known. Obviously, children do acquire both vocabulary and grammar. So, how do they accomplish these inter-related tasks?

Studies on children’s language development have investigated the intimate relation between vocabulary and grammar development. Vocabulary might provide the base on which grammatical knowledge is built (Caselli et al., 1999; Marchman & Bates, 1994; Szagun et al., 2006); alternatively, vocabulary and grammar may be inter-dependent (Dionne et al., 2003; Moyle et al., 2007), such that developing knowledge in one has a mutually beneficial effect on the other; or, alternatively, both vocabulary and grammar development may be driven by a general language learning ability (Brinchmann et al., 2019; Dixon & Marchman, 2007; Hoff et al., 2018; Language and Reading Research Consortium, 2015, 2018). Testing children’s language skills provides us with a measure that is a composite of children’s ability to learn language together with variation in the communicative environment to which children are exposed, the latter of which can have a substantial effect on language development (e.g., Dale et al., 2000; Dionne et al., 2003; Rowe, 2012). In order to isolate children’s learning skills from environmental contributions, actually testing this ability to learn may indicate more directly the mechanisms that underly children’s different language development trajectories.

Thus, instead of measuring children's natural language abilities – the outcome of children's learning skills – an alternative approach to understanding language development is to probe language learning skills themselves.

Artificial language learning tasks provide a possible means by which the mechanisms central to language learning can be interrogated. Statistical learning skills have been closely related to language development (Arciuli & Simpson, 2012; Kidd, 2012), and employing artificial languages has enabled researchers to highlight children's varied sensitivity to a range of statistical properties present in language, including both adjacent and non-adjacent dependencies between syllables in words, and words in phrases (Gomez & Gerken, 1999; Raviv & Arnon, 2018; Saffran, 2003). Non-adjacent dependencies are of particular theoretical interest in language learning because they underwrite phonotactic and morphological structure of words and they also occur in grammatical structures of sentences in natural language (Newport & Aslin, 2004), such as in number agreement between nouns and verbs (e.g., *the penguin that is grey likes fish*). The ability to identify non-adjacencies is thus an important aspect of word learning and grammatical processing (Wang, Zevin, & Mintz, 2019). In addition to identifying non-adjacencies, acquiring grammar requires ability to flexibly apply and generalise such non-adjacencies to new words occurring in a similar structure (e.g., *the penguin that waddles eats fish*) (Wilson et al., 2019).

In a landmark study, Peña et al. (2002) devised an artificial language in which processes involved in learning words and learning grammar could be assessed by testing adult participants in different ways from the same input, by determining participants' sensitivity to, and use of, non-adjacent dependencies. Peña et al.'s (2002) language incorporated non-adjacent dependencies that were taken to either define words, in a segmentation task, or define the grammatical structure of the language, requiring generalisation of the non-adjacency structure. Participants heard streams of continuous

syllables of the form AXB, where the A and B syllables reliably co-occurred, but the X syllable varied over three other syllables (e.g., *da-ro-pi-go-la-tu-da-la-pi-bu-fi-ko-go-ro-tu*... where in this example the A\_B non-adjacent pairs were *da\_pi*, *go\_tu*, and *bu\_ko*, and the X syllables were *ro*, *la*, and *fi*). After listening to the sequence, participants were tested on their ability to extract words from the language by comparing selection of a word as defined by the non-adjacent statistical structure of the language (e.g., *da-ro-pi*), from a part-word which occurred during the training speech but that spanned across the boundary between two non-adjacent dependencies (e.g., *ro-pi-go*). Participants were also tested on whether they were able to generalise to new sequences that respected the non-adjacent structure, by comparing selection of a generalised sequence (e.g., *da-go-pi*) from a part-word (e.g., *ro-pi-go*). Adult participants were able to segment the speech to distinguish words from part-words, but not able to distinguish generalised sequences from part-words. However, when short pauses were placed between the triplets, participants were able to generalise the grammar. Note that this generalisation task was more difficult than the segmentation (word learning) task, as segmentation requires recognition of a sequence consistent with the statistical structure of the language to which the learner has been exposed, whereas generalisation requires flexibly adjusting to novel sequences within the constraints of the structure. Consequently, accuracy is found to be lower for generalisation than segmentation for these tasks (Perruchet et al., 2004; Peña et al., 2002).

Frost and Monaghan (2016) repeated the studies with adults using Peña et al.'s (2002) artificial language with non-adjacent structure in order to determine whether there was a segregation of segmentation and generalisation for this task. They replicated the results showing accurate segmentation performance for these non-adjacencies. However, they also tested the acquisition of generalised grammar sequences in a different way to Peña et al. (2002). The generalised sequences in Peña et al.'s (2002) study interposed an A or a B

syllable from one non-adjacency within another non-adjacency (e.g., *da-go-pi* contained *go* which was an initial syllable in another sequence *go-X-tu*). In Frost and Monaghan (2016) a novel syllable that had not occurred during the listening phase of the study was used within the non-adjacency. This innovation avoided possible interference from syllables occurring in different positions in the test stimuli than had occurred during training, as in Peña et al.'s (2002) study, meaning that a more direct test of the grammatical learning could be accomplished. Frost and Monaghan (2016) showed that both segmentation and generalisation could be achieved from exactly the same speech input, and, in contrast to Peña et al.'s (2002) contention, that prior acquisition of words through segmentation was not required before generalisations of the structure were available to the learner. Frost and Monaghan (2017), using the same language to test segmentation and generalisation, found that sleep affected these two tasks differently, potentially indicating that, though both can be learned simultaneously, representation of words and grammar from the language may be distinct.

These studies were conducted with adults, but Marchetto and Bonatti (2013, 2015), using a slightly simplified version of Peña et al.'s (2002) non-adjacency artificial language, showed that infants too had the ability to segment words and generalise structure from artificial speech sequences. Marchetto and Bonatti (2015) found that 12-month-old children could detect words in continuous speech, and were able to generalise the grammar of the language if pauses were interspersed between the triplets. Whether these pauses are necessary for children in their learning is an issue we return to in the Discussion.

However, the value of these artificial language learning studies resides in their ability to relate to natural language processes, so the validity of these tasks with regard to natural language abilities must also be established. There is a growing body of research linking artificial language learning tasks with natural language. For instance, Cheung et al. (2022) found that whether children were late talkers or typically developing at 24 months predicted

their ability to retain words in an artificial word learning task at 40 months. Similarly, Ahufinger et al. (2021) and McGregor et al. (2022) found that children who were diagnosed with DLD demonstrated impaired learning in an artificial word learning task. In each of these studies, words were individuated in the speech, and only word learning was tested.

Relating early language learning ability in terms of both segmentation and generalisation skills would provide a further benefit over studies that have linked word learning to later language development, because this would also enable interactions between vocabulary and grammar to be examined. A step towards bridging multiple language learning processes detected through artificial language learning as predictors of natural language ability was taken by Frost et al. (2020). In their study, 17-month-old children's ability to learn an artificial language was related to natural language vocabulary skills from 19 through to 30 months. Frost et al. (2020) exposed children to the artificial language from Frost and Monaghan (2016), and then tested their segmentation and generalisation from this artificial language. Performance on both artificial language segmentation and generalisation were then examined in relation to children's vocabulary measured by a communicative development inventory completed when the children were aged 19, 21, 24, 25, 27, and 30 months. Children's segmentation in the artificial language was found to predict natural language vocabulary size at each time point. However, no significant relation was found between generalisation from the artificial grammar and the children's natural language vocabulary development.

Frost et al.'s (2020) study suggests that children's ability to use non-adjacent dependencies to segment words at 17 months was predictive of natural language vocabulary development over the next 13 months. The lack of a significant relation between artificial language structural generalisation and natural language vocabulary could be due to one of two possibilities. First, there may be a division between grammar and vocabulary in language

development, such that artificial language segmentation relates to natural language vocabulary development and artificial language generalisation relates to natural language grammar. In this case, we would predict that the artificial language generalisation task should relate to a measure of natural language grammar skills, and the artificial language segmentation task should continue to relate to measures of natural language vocabulary. Second, it may be that there is no clear division between vocabulary and grammar development (e.g., Brinchmann et al., 2019), and that children's ability to learn grammatical structure, as measured by the artificial language generalisation task, may be relevant for vocabulary learning but that this may manifest itself only later in children's language development.

There are a wide range of contributors to children's language learning that have changing influences as children's language develops. In terms of environmental variables, the quantity of children's early language experience is initially important for children's vocabulary development, but the richness of children's input, rather than the quantity, becomes more important as children grow older (Rowe, 2012). Children also become more sensitive to socio-pragmatic cues as they age, which are critical to supporting language learning (Bruner, 1983; Donnelly & Kidd, 2021). The types of information that children build on for learning from within the language also changes with age, with the syntactic constraints of words in sentences becoming more important in defining their meanings (Fisher et al., 2010; Naigles, 1996; Pinker, 1989). For example, Gleitman et al. (2005) suggested that early vocabulary can be acquired effectively by mapping words onto their referents in the environment, and so acquisition of the child's vocabulary in the early stages requires skills involving identifying the word in speech and linking it to a referent. However, as the vocabulary grows, acquisition of new words requires understanding their role in the sentence rather than just linking to concrete referents in the environment. Later vocabulary

development then becomes more dependent upon syntactic bootstrapping, whereby words are acquired by incorporating grammatical structure information to constrain and develop later lexical mappings (St Clair et al., 2010). Thus, at 30 months, natural language vocabulary development may be closely related to ability to identify words in speech, explaining the link between artificial language segmentation and vocabulary development (Frost et al., 2020). However, after 30 months, we may see greater involvement of grammatical learning abilities predicting more advanced stages of vocabulary learning. If this is the case then we would predict that the artificial language segmentation and generalisation tests should both relate to natural language vocabulary at later stages in language development.

In this study, we investigated the extent to which the artificial language learning tasks of segmentation and generalisation measured at 17 months relate to natural language vocabulary and grammar tasks at 54 months. These ages were selected because children at 17 months are close in age to the point at which children are first able to detect non-adjacent dependencies (Marchetto & Bonatti, 2015), and, for our sample, children at 54 months are just beginning formal schooling, which is an important age regarding children's language development and its influence on later academic achievement (von Hippel et al., 2017). We know that the artificial language segmentation at 17 months can predict natural language vocabulary ability up to 30 months (Frost et al., 2020). We also know that there is a close relationship between natural language vocabulary and grammar skills in preschool children (Brinchmann et al., 2019; Language and Reading Research Consortium, 2015). Our first research question was thus whether the artificial language segmentation task predicts natural language vocabulary development over a longer period, up to the point at which children commence formal schooling in the UK (at 54 months), and also, whether this artificial language segmentation task can predict children's grammar skills at 54 months. If so, language learning in a simple segmentation task early in children's development could be

useful to provide insight into processes involved in long-term language development across the preschool years.

Our second research question investigated whether the artificial language generalisation task at 17 months was useful for predicting vocabulary and grammar at 54 months. Though there is a correlation between natural language vocabulary and grammar skills (Brinchmann et al., 2019), and these skills can be learned at the same time (Frost & Monaghan, 2016), they may still be underwritten by different processes (Frost & Monaghan, 2017; Peña et al., 2002). If there is a distinction between vocabulary and grammar skills learning then we might observe that the artificial language generalisation task only relates to natural language grammar skills, and conversely that the artificial language segmentation task may relate only to natural language vocabulary skills. Alternatively, the generalisation task may not relate to natural language skills at all, consistent with Frost et al.'s (2020) finding that, at 30 months, no relationship was evident between artificial language generalisation and natural language vocabulary. A further possibility is that the greater complexity of the generalisation task may emerge as a valuable predictor of both vocabulary and grammar development at a later stage in children's language development (Gleitman et al., 2005). As natural language becomes more complex, the child's ability to build this complexity may be better predicted by more complex statistical processing required for generalising the structure of language, and utilise this structure to acquire new words.

In this study, we investigated the same cohort of children who had been tested on their vocabulary development at 30 months in Frost et al. (2020). We followed these children to age 54 months, at the point at which they were just beginning formal schooling in the UK, and then tested their natural language vocabulary and grammar skills on a number of different tasks. In a pre-registered analysis

([https://osf.io/8y67m/?view\\_only=ab7fbaf576e54fac9a2245ee4a6adf8b](https://osf.io/8y67m/?view_only=ab7fbaf576e54fac9a2245ee4a6adf8b)), we determined

whether the 17-month artificial language segmentation and generalisation learning tasks could predict language skills 37 months later.

## Method

### Participants

Participants were from a cohort of children from 95 families recruited as part of a larger longitudinal project in the North West of England, the Language 0-5 project (Rowland et al., unpub.). For more details of the project see the Language 0-5 Project OSF site (<https://osf.io/kau5f/>). Nine of these families had a history of language delay or dyslexia, but the children were typically developing when tested at 17 months in terms of their receptive vocabulary. All children were monolingual, born at term, and had normal vision and hearing at both testing timepoints. For the 17-month test of artificial language learning, there were 71 children tested (40 females, 31 males; aged between 16.5 and 17.5 months, mean age = 517 days). For the 54-month tests of natural language skills, there were between 62 and 72 children who completed each test.

### Artificial language learning task at 17 months

#### Materials

For the artificial language learning task, the speech was composed of four sequences where the first and third syllable (*ba\_so, li\_fe*) co-occurred, and the second syllable (*mu, ga*) could vary over one of two syllables. The four sequences that occurred were thus *bamuso, bagaso, limufe, and ligafe*. Speech stimuli were produced at 140Hz monotone pitch using the Festival speech synthesiser (Black et al., 1990).

For the training, sequences were synthesised as continuous speech such that co-articulation occurred between every syllable (within and across sequence boundaries). There were no direct repeats of sequences, and each of the six sequences occurred 200 times. All infants heard the same training sequence. The training speech lasted for approximately 15

minutes. Speech faded in for 5s at the beginning and out for 5s at the end of the training, and was presented through a loudspeaker attached to a computer.

For the segmentation testing trials, the four sequences consistent with the co-occurring non-adjacent syllables were used as the target stimuli, and foil sequences were created from sequences that straddled the boundary between the target sequences, so they occurred during the training but were not consistent with the statistical structure of the language. The foil sequences were *sobamu*, *feliga*, *gasoli*, and *mufeba*.

For the generalisation testing trials, the target stimuli were constructed from the non-adjacent dependent first and third syllables of the target sequences, with a novel syllable intervening between them. Thus, the generalisation targets were *baniso*, *baposo*, *lidufe*, and *livefe*. The foil stimuli used the last syllable of one of the training sequences and first syllable of another sequence with a novel syllable occurring before or after these syllable pairs. The four generalisation task foils were *posoba*, *nifeli*, *solive*, and *febadu*. Segmentation and generalisation testing trials were presented on a computer screen, with sound presented through loudspeakers.

### **Procedure**

All children were tested in exactly the same way for the artificial language learning tasks (with the same order of tasks and same order of stimuli within tasks). This reduced the inter-individual variation due to extraneous contributions and follows protocols required for adapting experimental tasks to individual difference measures (Cooper et al., 2017; Panter et al., 1992). However, this individual differences approach to reducing inter-individual variation introduces the possibility that there may be interference effects from one task to another. We return to this point in the Discussion.

During training, the child listened to the training speech for 15 minutes played at a comfortable volume while they engaged in quiet, non-verbal play with the experimenter.

Then, the child was tested using an adapted head-turn preference paradigm (Kemler Nelson et al., 1995). For testing, the child was moved to a car seat positioned in front of an SR Research Eyelink 1000+ eyetracker (SR Research: Ottawa, Ontario, Canada) at a distance of 580-620mm from a 17" computer screen. The eyetracker was operated in remote mode with the remote arm configuration and a target sticker which tolerated a small amount of movement by the child. Speakers were positioned to the left and right side of the screen.

Testing commenced with a five-point calibration. Then, the child was presented with the segmentation test items. Infants heard a test item repeated with a 500ms intervening pause, played through either the left or right speaker. The sounds were accompanied by an animated clip of a slow-moving hand against a black background, presented on the left or right of the screen corresponding to the sound source, similar to Marchetto and Bonatti's (2015) paradigm. Trials lasted up to 65s (as in Marchetto & Bonatti, 2013, 2015), but terminated when the child looked away from the slow-moving hand stimulus for more than 2s. A fixation stimulus then appeared at the centre of the screen, and the next trial was presented when the child fixated on this for 2s. Each target and foil stimulus was presented twice, so there were 16 trials in all.

After the segmentation test, the child watched a cartoon for 135s with non-verbal soundtrack, and then proceeded with the generalisation test. The procedure was identical to that of the segmentation test, and there were 16 trials in total.

### **Analysis and scoring**

For the artificial language learning tasks, responses were filtered by removing trials shorter than 700ms (which was the approximate duration of one stimulus) and trials longer than 2SD beyond the mean looking time across trials. Data from children who failed to provide responses for at least one target and foil trial were removed, and this resulted in data for one child being removed for the segmentation task, and data from nine children being

removed for the generalisation task. The mean number of trials included per child was 11 (range = 2 -16) for the segmentation task and 9 (range 2-14) for the generalisation task.

For both artificial language learning tests, we used the individual measure of preference for each child computed over the whole set of stimuli, developed by Frost et al. (2020). For the segmentation task, this measure takes the Cohen's *d* effect size of preference for looking to words over part-words for each individual child. A positive value for that child indicates longer looking to words over part-words, a negative value indicates longer looking to part-words over words, and a value close to zero indicates no preference. For the generalisation task, the measure was the effect size of each child's looking time to generalised sequences over part-words. A positive value indicated the child looks more to the generalised sequences than part-words, a negative value indicated more looks to the part-words, and a value close to zero indicates no preference.

## **Natural language tasks at 54 months**

### **Materials**

Natural language vocabulary was tested with a set of standardised assessments. We used the British Picture Vocabulary Scale (BPVS-3, Dunn et al., 2009) as a receptive vocabulary test, and the Renfrew Word Finding Vocabulary Test (Renfrew & Mitchell, 2010) as an expressive vocabulary test. Natural language grammar was tested with the Test for Reception of Grammar (TROG-2, Bishop, 2003) as a receptive grammar test, and the Renfrew Bus Story revised edition (Renfrew, 2010) as an expressive grammar test.

### **Procedure**

At 54 months, the children in the study were tested on a large battery of assessments, and four of these tests were included in the pre-registration for this study which focused on expressive and receptive tests of vocabulary and grammar. The tests that we analysed in the

current study were administered in the following order: (1) the TROG-2; (2) the Renfrew Word Finding Vocabulary Test; (3) the Renfrew Bus Story Task; and (4) the BPVS-3. For all assessments, children sat in a quiet room with the experimenter, accompanied by a caregiver. Children completed each assessment sitting opposite the experimenter at a table.

The TROG-2 was used to assess children's receptive grammar. Children were shown groups of four pictures and asked to point to the picture that matched the sentence spoken by the experimenter. The sentences increased in difficulty, from simple to complex as the test progressed. There were four sets of pictures per block and 20 blocks in total. We began at block A for all children, and proceeded until they reached the discontinue criterion for the test, which was five failed blocks in a row.

The Renfrew Bus Story Test was used to measure children's expressive grammar through narrative speech. The experimenter told children a story about a bus that was illustrated with 12 pictures. Following this, children were asked to retell the story. This was recorded with a Dictaphone and transcribed and coded after the session in CHAT format (MacWhinney, 2000) using the associated Computerized Language Analysis (CLAN) software.

The BPVS-3 was used to assess children's receptive vocabulary. Children were presented with an array of four pictures and were asked to select the picture that matched the word spoken by the experimenter. The words increased in difficulty throughout the assessment. There were 12 picture arrays in each set and 14 sets in total. At 54 months children started at set 2 and all items were presented until they reached the discontinue criterion in the test, which was eight or more wrong in one set.

The Renfrew Word Finding Vocabulary Test assessed children's expressive vocabulary. Children were shown individual line-drawn pictures and asked, 'What's this?'

for each picture. The pictures were presented in order of difficulty until the discontinue criterion was reached, which was failure to name five pictures in a row.

### **Analysis and scoring**

For the TROG-2, we used the raw scores for the test. For the Renfrew Bus Story task, we used two measures of expressive grammar from the transcriptions scored according to the manual (Renfrew, 1994). The first was sentence length score, which was calculated as the mean length in words of each child's 5 longest utterances. These scores were rounded to the nearest whole number. The second was subordinate clause score, where children were awarded a score of 1 if a sentence contained a subordinate clause, and 0 if not. The total number of sentences containing subordinate clauses was the resulting measure. Sentence length and subordinate clause use are likely to be related, as longer sentences are more likely to contain subordinate clauses, but subordinate clauses provide a more focused assessment of the hierarchical structural complexity of children's grammatical productions.

For the BPVS-3 we used the raw scores for productive vocabulary. For the Renfrew word finding vocabulary test, we again used the raw scores of accuracy.

## **Results**

All statistical analyses were pre-registered, and anonymised data are openly available here: [https://osf.io/8y67m/?view\\_only=ab7fbaf576e54fac9a2245ee4a6adf8b](https://osf.io/8y67m/?view_only=ab7fbaf576e54fac9a2245ee4a6adf8b).

### **Descriptive statistics**

Descriptive statistics for all the measures are shown in Table 1, and the correlations among the measures are shown in Table 2. Table 1 also reports the standardised scores for the BPVS-3 and TROG-2 tests.

Table 1. Descriptive statistics for all measures.

Measure	N	mean	sd	min	max
<b>Artificial Language Learning</b>					
Segmentation	70	-0.30	0.49	-1.41	0.76
Generalisation	61	0.36	0.74	-2.18	1.88
<b>Vocabulary Measures</b>					
BPVS-3	71	78.89	11.61	47	106
Renfrew Vocabulary	72	33.25	4.32	23	42
<b>Grammar Measures</b>					
TROG-2	72	7.10	3.70	1	17
Renfrew Subordinate Clause	62	0.98	1.11	0	4
Renfrew Sentence Length	62	8.43	2.40	4.40	13.80
<b>Standardised scores</b>					
BPVS-3	72	115.1	8.1	90	131
TROG-2	72	104.7	17.0	76	145

Table 2. Pairwise correlations among the measures.

	Artificial Language Generalisation	BPVS -3	Renfrew Vocabulary	Renfrew Subordinate Clause	Renfrew Sentence Length	TROG- 2
Artificial Language	-.07	-.09	-.15	.09	.17	-.03
Segmentation						
Artificial Language		-.31*	-.10	-.22	-.29*	-.14
Generalisation						
BPVS-3			.47***	.45***	.43***	.64***
Renfrew Vocabulary				.39**	.41***	.40***

Renfrew Subordinate	.78***	.30*
Clause		
Renfrew Sentence		.37**
Length		

---

Note: \*\*\*  $p < .001$ , \*\*  $p < .01$ , \*  $p < .05$ .

In order to analyse the relations among the 17-month artificial language learning measures and vocabulary and grammar abilities at 54 months, we conducted structural equation modelling, using lavaan in R (Rosseel, 2012). We used full information maximum likelihood to account for missing values in the dependent variables. For the modelling, we used standard fit indices with cut-off points as recommended by Hu and Bentler (1999), with good model fit indicated by root mean square error approximation (RMSEA)  $< .06$ , a comparative fit index (CFI)  $> .95$ , and Tucker-Lewis index (TLI)  $> .95$  (Hu & Bentler, 1999), and  $\chi^2$  difference tests, where a significant difference would indicate that the model and data diverge and that the model is therefore not a good approximation to the data.

### **Confirmatory factor analysis: Determining relations among the natural language measures**

In accordance with our pre-registered analysis plan, we first determined whether the two vocabulary measures (BPVS-3 and Renfrew Vocabulary) and the three grammar measures (TROG-2, Renfrew sentence length, Renfrew subordinate clause score) could be effectively combined into two latent variables (one for vocabulary and one for grammar) using confirmatory factor analysis. The model did not provide a good fit to the data, RMSEA = .290, CFI = .806, TLI = .515,  $\chi^2(4) = 28.144$ ,  $p < .001$ . Furthermore, the model fit resulted in a correlation that problematically exceeded 1 between the vocabulary and grammar factors (indicating that the covariance matrix was not positive definite, Chen et al., 2001). This was

likely due to the high correlation between the TROG-2 and the vocabulary measures (see Table 2).

We followed our pre-registration plan by next examining modification indices, which test how model fit to the data changes by relaxing a constraint of the model (i.e., adding a path into the model). We found the greatest change in fit by adding TROG-2 to the vocabulary factor, indicating that the TROG-2 measure shared variance with the vocabulary measures, as well as the grammar measures. The correlation between the TROG-2 and the BPVS-3 measures may be partly due to shared similarities in terms of the way in which the task is presented and how children respond: both are receptive tests, requiring the child to listen to speech, then select one of four pictures. LARRC, Logan, and Jia (2018) found a similar close relationship between these tests for children in the same age range. We thus addressed this poor fit of the model to the data by conceptualising the natural language measures as a single factor – that is, a general language skill construct representing all the natural language vocabulary and grammar measures together – rather than separating vocabulary and grammar into separate factors, to see if this resulted in a good fit to the data.

Running this confirmatory factor analysis resulted in poor fit to the data, RMSEA = .280, CFI = 0.773, TLI = 0.545,  $\chi^2(5) = 33.324$ ,  $p < .001$ . Again, according to our pre-registration plan, we tested the modification indices, which indicated that adding covariance between the Renfrew sentence length and Renfrew subordinate clause measures would improve model fit, and so we also added this pathway. As these measures were computed from the same task, it is likely that they do covary. Adding this covariate resulted in a good fit, RMSEA = .000, CFI = 1.000, TLI = 1.012,  $\chi^2(4) = 3.381$ ,  $p = .496$ .

**Structural Equation Model: Linking artificial language learning at 17 months to language skills at 54 months**

We next applied these results of the confirmatory factor analysis to structural equation modelling. First, we conducted the pre-registered original plan for the structural equation model relating the 17-month measures of artificial language segmentation and generalisation to the vocabulary and grammar factors. We anticipated that this structural equation model would not provide a good fit, because of the variance shared between the TROG-2 measure and the BPVS-3 vocabulary measure, which is not respected in the model, and this was confirmed to be the case. These results are reported in Supplementary Information S1.

The second way we addressed structural equation model fit, paralleling the approach taken in the confirmatory factor analyses, was to construct a structural equation model with one latent variable – a general language construct – to represent shared variance among all the language measures at 54 months, and including covariance between the Renfrew sentence length and Renfrew subordinate clause measures. This model enabled us to test whether the natural language measures at 54 months were fit well by conceiving of them as a unitary skill, and then determine whether the 17-month artificial language learning measures related to this general language skills latent variable at 54 months. The general language construct model, however, does not enable us to distinguish the extent to which the artificial language learning measures predict natural vocabulary and grammar development individually, but it does enable us to determine whether the shared variance between natural language vocabulary and grammar skills is predicted by either artificial language segmentation or generalisation performance. Interrogating the correlations in Table 2 also provides us with insight about relations among particular natural language measures and the artificial language learning tasks.

The model included one latent variable – a general language skills variable – and covariance between the Renfrew sentence length and subordinate clause scores. The model fit was good, RMSEA = .043, CFI = .987, TLI = .978,  $\chi^2(12) = 13.326$ ,  $p = .346$ . This model is

shown in Figure 1. The model demonstrates a significant negative relation between the artificial language generalisation task at 17 months and the language measures at 54 months, indicating that longer looking times to foil stimuli over target generalisation stimuli related to higher natural language scores. Taken together with the correlations among the individual natural language measures, the results demonstrate that the artificial language segmentation task at 17 months did not predict either natural language vocabulary nor natural language grammar skills at 54 months, but that the artificial language generalisation task did predict a combined measure of natural language vocabulary and grammar skills.

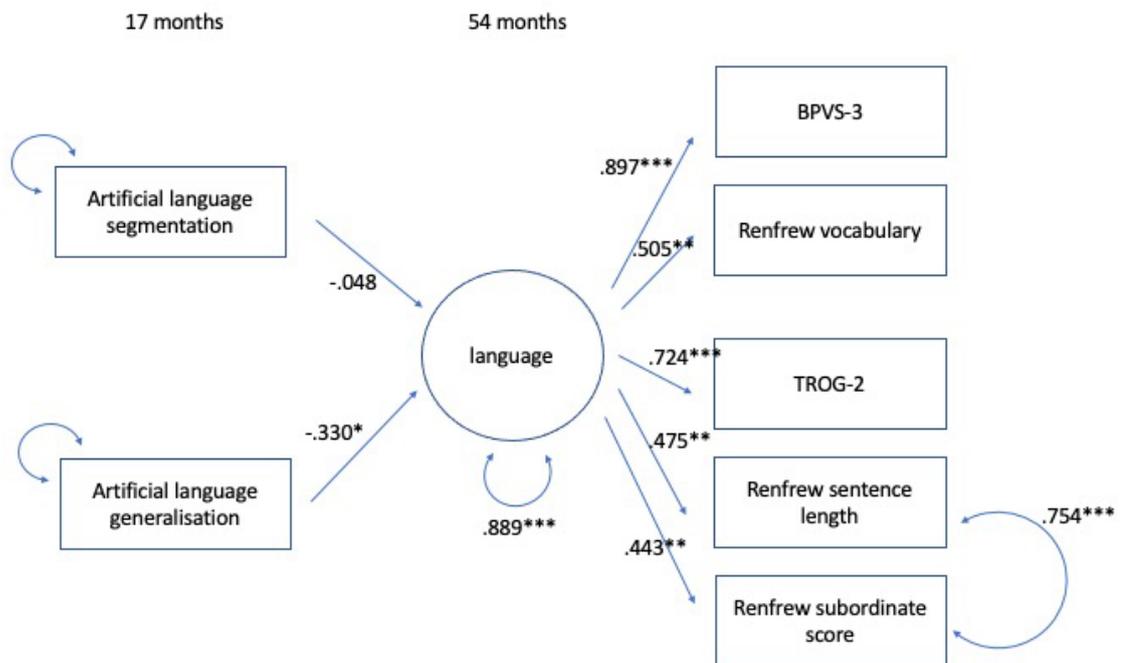


Figure 1. Relations between 17-month tests of artificial language segmentation and generalisation and general natural language skills at 54 months. Values on paths are standardised estimates. \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ .

A

B

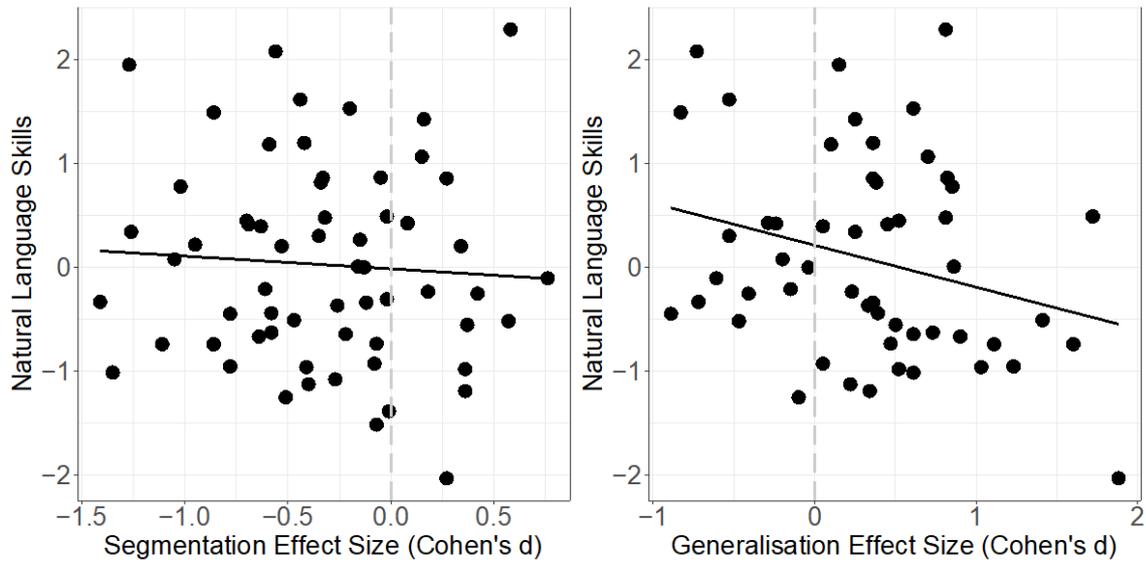


Figure 2. Individual children's scores on (a) artificial language segmentation and (b) artificial language generalisation tasks compared to the combined natural language vocabulary and grammar latent variable from the confirmatory factor analysis model. Notes: lines show linear regression of the relation, with grey shaded areas indicating 95% CI; for the generalisation task, one outlier on the generalisation task was removed from the figure to show the relation more clearly.

The relations between artificial language segmentation and generalisation with the combined natural language latent variable are shown in Figure 2.

## Discussion

Our primary aims for this study were to determine whether and how measures of segmentation and generalisation in an artificial language learning task, presented to children at 17 months, could predict natural language vocabulary and grammar performance 37 months later, just prior to the children beginning school. Language and communication skills at school entry are a key predictor not only of literacy development but also academic success more generally (von Hippel et al., 2017). Understanding the mechanisms underlying

children's language skill development early in childhood has key advantages for supporting children who are likely to be delayed in terms of their communicative development (Dale et al., 2003; Leonard, 2009, 2014; Reilly et al., 2010; Rescorla, 2009).

Regarding our first research question to determine the link between artificial language segmentation at 17 months and language development at 54 months, we assessed a structural equation model which represented a combined language variable, constructed from a set of natural language vocabulary and grammar skill tasks given to children at 54 months. We found no evidence of segmentation performance at 17 months predicting vocabulary or grammar 37 months later. In previous analyses of the same children's language development up to 30 months, Frost et al. (2020) found that the 17-month segmentation task did predict vocabulary: children who showed a greater preference for foils (non-conforming stimuli) over words (stimuli that conformed to the non-adjacent statistical language structure) had larger vocabularies at age 30 months than children who showed a greater preference for words over foils.

So, what changed between 30 months and 54 months in terms of vocabulary development? One possibility is that segmentation of an artificial language vocabulary learning task can account for language development in earlier stages, when the structures that the child is required to learn are less complex, and the child's language mastery is less complete. Consistent with Frost et al. (2020), Weyers et al. (2022) found that 8 to 10 month old children with better language skills showed sensitivity to non-adjacent sequences in an artificial language as measured using EEG, but no sensitivity was detected in children with poorer language skills. Similarly, Gerbrand et al. (2022) found that visual sequence processing (of adjacent dependencies) at 10 and 18 months predicted children's vocabulary at 18 months. Thus, statistical processing of sequences, required to identify words in speech, may effectively predict children's earlier vocabulary skills. However, the ability to segment

words may become less relevant to later stages of natural language vocabulary development. Though there are many changes in children's sensitivity to, and use of, different information sources in language learning (Bruner, 1983; Donnelly & Kidd, 2021; Pinker, 1989; Rowe, 2012), the isolation of mechanisms associated with segmentation of speech into words, and generalisation of structure from the non-adjacent statistical dependencies in speech, is consistent with accounts of language development that describe the growing importance of the role of syntactic constraints on language learning (e.g., Fisher et al., 2010). For example, Gleitman et al. (2005) proposed that after the initial stages of vocabulary development, the structural relations among words in speech become more important in vocabulary acquisition, whereas at earlier stages of vocabulary development word learning can proceed by merely identifying the speech sounds and linking them to referents. Our artificial language task thus relates to Gleitman et al.'s (2005) hypothesis: segmentation of words in artificial language speech is predictive of vocabulary development up to 30 months, but not at 54 months. Though syntactic bootstrapping accounts of language development are often discussed with regard to verb learning, using the contextual constraints of words is known to also be helpful for acquisition of words from other grammatical categories (e.g., Mintz, 2002; Monaghan et al., 2007). Furthermore, processing the sentential context of a word becomes more important as children must learn different senses of words, and determine how relational terms affect the meaning of the content words in an utterance, which are processes that increase in importance as children's language progresses over the first 5 years of life (e.g., Harris, Golinkoff, & Hirsh-Pasek, 2011; Naigles, 1996). The computation of context, therefore, likely applies accumulatively to words from all grammatical categories.

An alternative possibility for the change in results from 30 months to 54 months is that the way in which vocabulary was measured affected the predictive relationship between the artificial language segmentation task and the natural language measures. In the measures

up to 30 months, children's vocabulary was estimated using Communicative Development Inventories (Alcock et al., 2020; Meints & Fletcher, 2001). These are caregiver report questionnaires that ask which words on a checklist a child can say and/or understand. In the measures given at 54 months, the child's vocabulary was directly interrogated either by measuring children's comprehension of spoken words by indicating one of a series of responses (for the BPVS-3), or by analysing the child's production of words in response to pictures (for the Renfrew Vocabulary test). However, it seems unlikely that there is a qualitative difference between direct tests of children's vocabulary and caregiver report measures, as CDI scores and direct measures of receptive vocabulary tend to correlate highly (Fenson et al., 1993; Pan et al., 2004).

Our second research question concerned whether children's ability to generalise over the structure of an artificial language at 17 months predicted later language development. Here, we did observe a significant effect. The structural equation modelling demonstrated that the artificial language generalisation task at 17 months predicted differences in natural language skills 37 months later. The children with higher vocabulary skills showed longer looking times towards novel foil stimuli (that did not conform to the grammatical structure of the language) over target stimuli (that did conform to the grammatical structure of the language). This is a similar direction as the relationship that Frost et al. (2020) observed for predicting vocabulary development up to 30 months from behaviour in the artificial word learning task at 17 months. In that study, children with a preference for looking longer at novel foil stimuli, that do not conform to the non-adjacent structure of the language (i.e., that were not words in the language) over target stimuli where the sequences do conform to the non-adjacent structure (that were words in the language), are those with larger vocabularies. In the current study, this novelty preference for non-conforming over conforming generalisation sequences – both containing a novel syllable that the children had not heard

before – related to their language abilities at 54 months. Note though that Frost et al. (2020) did not find a relation between artificial language generalisation performance and vocabulary up to 30 months.

In terms of evidence for distinctions between vocabulary and grammar skills in early natural language development, the results are inconclusive. Our study was not specifically designed to test how these natural language skills develop, which would require regular measures of vocabulary and grammar longitudinally and investigating the relation between those skills (e.g., Brinchmann et al., 2019; Hoff et al., 2018). Rather the focus of our study was the cognitive precursors of those natural language vocabulary and grammar skills. What we did find is that a confirmatory factor analysis with vocabulary and grammar as separate latent variables did not provide a good fit to the data, because in our data there was a strong relationship between the vocabulary and the grammar measures: children whose vocabulary skills were strong also demonstrated good grammar skills. In part due to this strong relationship, the model with one latent variable representing a general language construct provided a good fit to the data, indicating that, for the set of tasks included in this study, there was substantial shared variance between vocabulary and grammar in children's language skills. The correlations shown in Table 2 indicate that there are correlations among tasks measuring grammar, and between the two measures of vocabulary, but that there are also correlations across the vocabulary and grammar tasks, which may also partially reflect the ways in which the skills were measured (such as the high correlation between the BPVS-3 measure of vocabulary and the TROG-2 measure of grammar which are both receptive measures requiring selection of one picture from a set).

These results are neither inconsistent with models of language development that posit a general cognitive skill that underlies development of both vocabulary and grammar (e.g., Brinchmann et al., 2019) nor also models that propose there is a degree of separation of these

skills (Caselli et al., 1999; Hoff et al., 2018; Marchman & Bates, 1994; Szagun et al., 2006). From our natural language data measures, deciding between these alternatives is not possible. Our study was designed, however, to relate the artificial language learning tasks to natural language measures, and there we see an emerging distinction in the learning abilities relating to language development at 54 months, with generalisation predicting language skills but segmentation not predicting language at this stage of development.

However, the set of natural language tasks measuring vocabulary and grammar spanned both receptive and expressive tasks, and as previously mentioned the correlation in performance between the receptive tasks, in particular, seemed to be strong. In the LARRC et al. (2018) study on 60-month-old children, there was a high correlation between vocabulary and grammar constructs ( $r = .935$ ), and, as in the current study, BPVS and TROG tests were included in the set of measures designed to assess vocabulary and grammar in natural language. In the Language and Reading Research Consortium (2015) study on children in pre-school and first year of formal schooling, vocabulary and grammar were found to not be reliably distinct constructs consistent with the current study. It may be that more reliable distinctions between vocabulary and grammar constructs emerge later in language development, beyond the preschool years, at least for receptive language.

In parallel with the debate about distinctions in vocabulary and grammar in children's language development, the artificial language learning tasks used at 17 months were originally devised to determine whether segmentation and generalisation are dissociable processes in language learning, highlighting distinctions in the mechanisms involved in word and grammatical learning. Peña et al. (2002) and Marchetto and Bonatti (2015) found that only when pauses were placed between sequences in the speech were learners able to generalise the structure of the language. Frost and Monaghan (2016) found that both segmentation and generalisation were possible when novel, rather than repositioned, syllables

were used in the test sequences, casting doubt on a computational distinction between segmentation and generalisation of this language at least for adult learners. The current study shows that pauses in the stimuli are not necessary for adjusting learning from segmentation of words to grammatical generalisation for infants either, as the individual differences on the segmentation *and* generalisation tasks (using a language without pauses between non-adjacent sequences) were both found to relate to language development (in Frost et al., 2020, and in the current study, respectively). Hence, children's identification of words and generalisation over the artificial language were both assessing early mechanisms relevant to language development. However, our study does point toward a dissociation in these two tasks in terms of how they relate to different stages of children's language development. At age 30 months, segmentation is predictive of vocabulary skills, whereas by 54 months, it is generalisation that predicts variance in language skills. Segmentation may be more critical early in language development, with generalisation becoming more important as the child's language matures and as the language experience and structures from which children learn become more complex (Gleitman et al., 2005; Huttenlocher et al., 2002; Theakston & Lieven, 2017).

A further advantage of using continuous speech, rather than presenting pauses between the sequences, is that we can then be more certain that the non-adjacencies are being responded to by the learners, rather than positional information. When sequences are separated by pauses, the learner may become sensitive first of all to the position of syllables within the sequence. Indeed, in a study with adults using the same language as Pena et al. (2002) with pauses between sequences, Perruchet et al. (2004) showed that learners had a preference for sequences that respected the position of syllables over part-words, where the syllable positions were consistent with those used in the words in training, but the non-adjacencies were violated. Hence, in studies with pauses, positional information encoding is

a potential confound with non-adjacent dependency processing. In a series of studies, Endress and Bonatti (2007) showed that adult learners first acquired syllable positions of non-adjacent triplets when there were pauses between the sequences, before acquiring dependencies among those syllables. However, when pauses were not present between the sequences, the positional information was not acquired before sensitivity to the non-adjacent dependencies, and so the positional information is unlikely to be a confound to determining processing of non-adjacencies when pauses, or other cues indicating positions of syllables, are not present within the sequences. Further, the use of novel syllables to test generalisation ensured that violations of syllables in particular positions were not driving the results, which could have been a further potential confound in previous studies where generalisation was tested by moving syllables from other positions in the sequences (see Frost & Monaghan, 2016, for discussion).

A potential issue in the design of the study was that the artificial language generalisation task occurred after the segmentation task for all children. This order effect may have manifested itself in two ways. First, it may be that those children that performed well on the second task were those more able to attend to tasks for longer, relating to general cognitive resources enabling endurance of an additional task. Second, it may be that there was contamination from the first task to the second, meaning the measure of generalisation is not distinct from learning that could have occurred during the segmentation testing.

We contend that these explanations are unlikely for a number of reasons. In terms of generalisation performance being a proxy for the cognitive resources of the child, it is possible that children with greater resources could attend to the task longer, thus showing more sensitivity to the generalisation stimuli, and children with more cognitive resources may have higher natural language skills (independent of the artificial language task). However, critically, in the generalisation task children showed sensitivity to the stimuli in two ways,

and this sensitivity related either to higher or lower language skills. Children who responded with novelty preference to the generalisation stimuli were those with higher language skills, whereas those with familiarity preference to the generalisation stimuli were those with lower language skills (see Figure 2b). Children who did not demonstrate sensitivity to the stimuli were those with intermediate language skills.

To test this effect further, we included an additional measure of cognitive resources – working memory – into the structural equation model to determine whether the artificial language tasks predicted natural language skills when working memory was also included as a predictor of natural language. The working memory test was the BRIEF-P digit span task (Gioia et al., 2000), which was measured with the children in this study at 37, 43, and 49 months. We selected the 49 month test due to it being the closest in time to the natural language measures taken at 54 months. As this analysis was not part of our preregistration, the results are reported in the Supplementary Information S2. The key results in the structural equation model shown in Figure 1 were unchanged: when working memory resources were also taken into account, artificial language generalisation, but not segmentation, still predicted natural language skills.

In terms of the second potential limitation – that generalisation performance could have been affected by segmentation performance – again, we suggest that this order effect does not seem to exert an effect on the key results. We attempted to minimise the possibility of learning and transfer from the stimuli used for testing segmentation to those used for testing generalisation in the design of the study. We placed a 2 minute pause between these tasks to mitigate against interference. In addition, as Frost et al. (2020) note, the segmentation test contains target stimuli containing non-adjacent dependencies, but also an equal number of foil stimuli containing no non-adjacent dependencies, thus reinforcing words and part-words to an equal degree. We do acknowledge, however, that there may have been greater

similarity between the segmentation and generalisation targets than between the foils for these tasks. However, if there was an influence of one task on the other then we would expect a correlation between the tasks, yet there was no significant correlation between segmentation and generalisation scores (see Table 2), meaning that sensitivity to the structure for segmentation testing did not result in higher or lower sensitivity for the generalisation test. Similarly, if ability to remember the segmentation test stimuli influenced performance on the generalisation test, then we would anticipate an effect of working memory on generalisation test performance. There was no significant correlation,  $r(55) = -.25, p = .089$ .

For these reasons, we are confident that the generalisation test scores are a reflection of the child's use of non-adjacent dependencies to generalise to novel, but conforming, stimuli.

Though children's communicative development is in large part a consequence of their environment (Dale et al., 2000; Dionne et al., 2003), there is still a component of individual differences in cognitive processes that affects development. The current study contributes to a growing literature that is uncovering the mechanisms involved in children's language development (Ahufinger et al., 2021; Cheung et al., 2022; Hartley et al., 2020; McGregor et al., 2022; Peter et al., 2019). The variance explained by the artificial language generalisation task is significant, but relatively small, accounting for approximately 11% of the variance in later natural language skills, and, so the value of our result is primarily to elucidate the processes involved in language development rather than to be useful as a predictor of natural language skills. Nevertheless, our direct measures of language learning may explain additional variance in language development alongside established measures such as early vocabulary size, nonword repetition performance, socioeconomic status, family history of language disorders, and gender (Dale et al., 2003; Hammer et al. 2017; Hodges et al., 2017; Law & Roy, 2008; Marini et al., 2017; Stokes & Klee, 2009; Rescorla, 2009).

In this study, we have investigated whether a language learning task administered early in children's language development – at 17 months – is able to predict language skills 37 months later. We found that it is. The results demonstrate the possibility of using laboratory-based learning tasks to probe children's ability to acquire language, to pinpoint different cognitive mechanisms involved in children's language learning, and predict the pattern of children's language development over the critical first few years of life.

## References

- Ahufinger, N., Guerra, E., Ferinu, L., Andreu, L., & Sanz-Torrent, M. (2021). Cross-situational statistical learning in children with developmental language disorder. *Language, Cognition and Neuroscience*, 36(9), 1180–1200.  
<https://doi.org/10.1080/23273798.2021.1922723>
- Alcock, K. J., Meints, K., & Rowland, C. F. (2020). *The UK communicative development inventories: Words and gestures*. Guilford, UK: J&R Press Ltd.<sup>[1]</sup><sub>[SEP]</sub>
- Arciuli, J., & Simpson, I.C. (2011). Statistical learning in typically developing children: The role of age and speed of stimulus presentation. *Developmental Science*, 14, 464–473.
- Bates, E., Bretherton, I., & Snyder, L. (1988). *From first words to grammar: Individual differences and dissociable mechanisms*. New York: Cambridge University Press.
- Black, A. W., Taylor, P., & Caley, R. (1990). *The festival speech synthesis system*. Centre for Speech Technology Research (CSTR), University of Edinburgh, UK. Retrieved from <http://www.cstr.ed.ac.uk/projects/festival.html>.
- Bishop, D. (2003). *Test for reception of grammar, version 2 (TROG-2)*. London, UK: Pearson Assessment.
- Brinchmann, E. I., Braeken, J., & Lyster, S. A. H. (2019). Is there a direct relation between the development of vocabulary and grammar? *Developmental Science*, 22(1), e12709.
- Bruner, J. (1983). *Child's talk: Learning to use language*. New York, NY: W.W. Norton.
- Caselli, C., Casadio, P., & Bates, E. (1999). A comparison of the transition from first words to grammar in English and Italian. *Journal of Child Language*, 26(1), 69-111.
- Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, 29(4), 468-508.

- Cheung, R.W., Hartley, C., & Monaghan, P. (2022). Multiple mechanisms of word learning in late talking children: A longitudinal study. *Journal of Speech, Language and Hearing Research, 65*(8), 2978-2995.
- Childers, J. B., Heard, M. E., Ring, K., Pai, A., & Sallquist, J. (2012). Children use different cues to guide noun and verb extensions. *Language Learning and Development, 8*(3), 233–254. <https://doi.org/10.1080/15475441.2011.585285>
- Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in Psychology, 8*, 1482. <https://doi.org/10.3389/fpsyg.2017.01482> [L<sup>1</sup>SEP]
- Dale, P. S., Dionne, G., Eley, T. C., & Plomin, R. (2000). Lexical and grammatical development: A behavioural genetic perspective. *Journal of Child Language, 27*(3), 619- 642. <http://dx.doi.org/10.1017/S0305000900004281>
- Dale, P. S., Price, T. S., Bishop, D. V. M., & Plomin, R. (2003). Outcomes of early language delay: I. Predicting persistent and transient language difficulties at 3 and 4 years. *Journal of Speech, Language, and Hearing Research, 46*(3), 544–560.
- Dionne, G., Dale, P., Boivin, M., & Plomin, R. (2003). Genetic evidence for bidirectional effects of early lexical and grammatical development. *Child Development, 74*, 394–412.
- Dixon, J. A., & Marchman, V. A. (2007). Grammar and the lexicon: Developmental ordering in language acquisition. *Child Development, 78*(1), 190-212. <http://dx.doi.org/10.1111/j.1467-8624.2007.00992.x>
- Donnelly, S., & Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development, 92*(2), 609-625.

- Dunn, L.M., Dunn, D.M., Sewel, J., & Styles, B. (2009). *The British Picture Vocabulary Scale (BPVS-3)*. 3<sup>rd</sup> edition. London: GL Assessment Ltd.
- Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., Pethick, S., & Reilly, J. S. (1993). *The MacArthur Communicative Development Inventories: user's guide and technical manual*. San Diego: Singular Publishing Group.
- Fisher, C., Gertner, Y., Scott, R. M., & Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(2), 143-149.
- Frost, R. L., Jessop, A., Durrant, S., Peter, M. S., Bidgood, A., Pine, J. M., Rowland, C.F., & Monaghan, P. (2020). Non-adjacent dependency learning in infancy, and its link to language development. *Cognitive Psychology*, 120, 101291.
- Frost, R. L. A. & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non- adjacent dependencies from continuous speech. *Cognition*, 147, 70-74.  
Doi:10.1016/j.cognition.2015.11.010
- Frost, R. L. A. & Monaghan, P. (2017). Sleep-driven computations in speech processing. *PloS one*, 12(1), e0169538.
- Gerbrand, A., Gredebäck, G., Hedenius, M., Forsman, L., & Lindskog, M. (2022). Statistical learning in infancy predicts vocabulary size in toddlerhood. *Infancy*, 27(4), 700-719.
- Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1(1), 3–55. [https://doi.org/10.1207/s15327817la0101\\_2](https://doi.org/10.1207/s15327817la0101_2).
- Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, 1(1), 23–64. [https://doi.org/10.1207/s15473341lld0101\\_4](https://doi.org/10.1207/s15473341lld0101_4).
- Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). Test review behavior rating inventory of executive function. *Child Neuropsychology*, 6(3), 235-238.

- Gómez, R.L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109–135.
- Hammer, C. S., Morgan, P., Farkas, G., Hillemeier, M., Bitetti, D., & Maczuga, S. (2017). Late talkers: A population-based study of risk factors and school readiness consequences. *Journal of Speech, Language, and Hearing Research*, *60*(3), 607–626.  
[https://doi.org/10.1044/2016\\_JSLHR-L-15-0417](https://doi.org/10.1044/2016_JSLHR-L-15-0417)
- Harris, J., Golinkoff, R. M., & Hirsh-Pasek, K. (2011). Lessons from the crib for the classroom: How children really learn vocabulary. In Neuman, S.B., & Dickinson, D.K. (Eds.), *Handbook of early literacy research: Volume 3*, 49-65. New York: Guilford Press.
- Hartley, C., Bird, L. A., & Monaghan, P. (2020). Comparing cross-situational word learning, retention, and generalisation in children with autism and typical development. *Cognition*, *200*, 104265.
- Hodges, R., Munro, N., Baker, E., McGregor, K., & Heard, R. (2017). The Monosyllable Imitation Test for toddlers: Influence of stimulus characteristics on imitation, compliance and diagnostic accuracy. *International Journal of Language & Communication Disorders*, *52*(1), 30–45.
- Hoff, E., Quinn, J., & Giguere, D. (2018). What explains the correlation between growth in vocabulary and grammar? New evidence from latent change score analyses of simultaneous bilingual development. *Developmental Science*, *21*, e12536.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.  
<https://doi.org/10.1080/10705519909540118>

- Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology*, *45*(3), 337-374.
- Kemler Nelson, D. G., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. A. (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development*, *18*, 111–116. [https://doi.org/10.1016/0163-6383\(95\)90012-8](https://doi.org/10.1016/0163-6383(95)90012-8).
- Kidd, E. (2012). Implicit statistical learning is directly associated with the acquisition of syntax. *Developmental Psychology*, *48*, 171–184.
- Language and Reading Research Consortium (2015). The dimensionality of language ability in young children. *Child Development*, *86*(6), 1948-1965.
- Language and Reading Research Consortium, Logan, J., & Jia, R. (2018). Modelling the nature of grammar and vocabulary trajectories from pre-Kindergarten to third grade. *Journal of Speech, Language, and Hearing Research*, *61*, 910-923.
- Law, J., & Roy, P. (2008). Parental report of infant language skills: A review of the development and application of the Communicative Development Inventories. *Child and Adolescent Mental Health*, *13*(4), 198–206. <https://doi.org/10.1111/j.1475-3588.2008.00503.x>
- Leonard, L. B. (2009). Is Expressive Language Disorder an accurate diagnostic category? *American Journal of Speech-Language Pathology*, *18*(2), 115–123. [https://doi.org/10.1044/1058-0360\(2008/08-0064\)](https://doi.org/10.1044/1058-0360(2008/08-0064))
- Leonard, L. B. (2014). *Children with Specific Language Impairment*. Cambridge, MA: MIT Press.
- Marchetto, E. & Bonatti, L. L. (2013). Words and possible words in early language acquisition. *Cognitive Psychology*, *67*(3), 130–150. Doi:10.1016/j.cogpsych.2013.08.001

- Marchetto, E. & Bonatti, L. L. (2015). Finding words and word structure in artificial speech: the development of infants' sensitivity to morphosyntactic regularities. *Journal of Child Language*, 42(4), 873-902. Doi:10.1017/S0305000914000452
- Marchman, V., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language*, 21, 339–366.
- Marini, A., Ruffino, M., Sali, M. E., & Molteni, M. (2017). The role of phonological working memory and environmental factors in lexical development in Italian-speaking late talkers: A one-year follow-up study. *Journal of Speech, Language & Hearing Research*, 60(12), 3462–3473.
- McGregor, K. K., Smolak, E., Jones, M., Oleson, J., Eden, N., Arbisi-Kelm, T., & Pomper, R. (2022). What children with Developmental Language Disorder teach us about cross-situational word learning. *Cognitive Science*, 46(2), e13094.
- Meints, K., & Fletcher, K. L. (2001). *Toddler communicative development inventory, a UK adaptation of the MacArthur-Bates communicative development inventories*. University of Lincoln.
- Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1), 91-117.
- Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, 55(4), 259-305.
- Moyle, M. J., Weismer, S. E., Evans, J. L., & Lindstrom, M. J. (2007). Longitudinal relationships between lexical and grammatical development in typical and late-talking children. *Journal of Speech, Language, and Hearing Research*, 50(2), 508–528.
- Naigles, L. R. (1996). The use of multiple frames in verb learning via syntactic bootstrapping. *Cognition*, 58(2), 221-251.

- Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48(2), 127–162.
- Pan, B. A., Rowe, M. L., Spier, E., & Tamis-LeMonda, C. (2004). Measuring productive vocabulary of toddlers in low-income families: Concurrent and predictive validity of three sources of data. *Journal of Child Language*, 31(3), 587-608.
- Panter, A. T., Tanaka, J. S., & Wellens, T. R. (1992). The psychometrics of order effects. In N. Schwarz, & S. Sudman (Eds.), *Context Effects in Social and Psychological Research*. New York, NY: Springer.
- Peña, M., Bonatti, L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, 298, 604-607. Doi :10.1126/science.1072901
- Perruchet, P., Tyler, M. D., Galland, N., & Peerean, R. (2004). Learning non-adjacent dependencies: No need for algebraic-like computations. *Journal of Experimental Psychology: General*, 133(4), 573–583.
- Peter, M., Durrant, S., Jessop, A., Bidgood, A., Pine, J., & Rowland, C. F. (2019). Does speed of processing or vocabulary size predict later language growth in toddlers? *Cognitive Psychology*, 115, 101238.
- Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT press.
- Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science*, 21(4), e12593. <https://doi.org/10.1111/desc.12593>
- Reilly, S., Wake, M., Ukoumunne, O. C., Bavin, E., Prior, M., Cini, E., Conway, L., Eadie, P., & Bretherton, L. (2010). Predicting language outcomes at 4 years of age: Findings from Early Language in Victoria Study. *Pediatrics*, peds.2010-0254. <https://doi.org/10.1542/peds.2010-0254>

- Renfrew, C. E. (2010). *Bus story test: A test of narrative speech*. Bicester: Speechmark.
- Renfrew, C. E., & Mitchell, P. (2010). *Word finding vocabulary test*. Bicester: Speechmark.
- Rescorla, L. (2009). Age 17 language and reading outcomes in late-talking toddlers: Support for a dimensional perspective on language delay. *Journal of Speech, Language, and Hearing Research, 52*(1), 16–30. [https://doi.org/10.1044/1092-4388\(2008/07-0171\)](https://doi.org/10.1044/1092-4388(2008/07-0171))
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36.
- Rowe, M. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development, 83*, 1762–1774.
- Rowland, C.F., Bidgood, A., Durrant, S., Peter, M., & Pine, J. M. (unpub.). *The Language 0-5 Project*. Unpublished manuscript, University of Liverpool. Available from <https://osf.io/kau5f/>.
- Saffran, J.R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science, 12*, 110–114.
- St. Clair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition, 116*(3), 341-360.
- Stokes, S. F., & Klee, T. (2009). The diagnostic accuracy of a new test of early nonword repetition for differentiating late talking and typically developing children. *Journal of Speech, Language, And Hearing Research, 52*(4), 872–882.
- Szagun, G., Steinbrink, C., Franik, M., & Stumper, B. (2006). Development of vocabulary and grammar in young German-speaking children assessed with a German language development inventory. *First Language, 26*(3), 259-280.  
<http://dx.doi.org/10.1177/0142723706056475>

- Theakston, A., & Lieven, E. (2017). Multiunit sequences in first language acquisition. *Topics in Cognitive Science*, 9(3), 588-603.
- von Hippel, P., Workman, J., & Downey, D. B. (2017). Inequality in reading and math skills comes mainly from early childhood: A replication, and partial correction, of ‘Are schools the great equalizer?’ *SSRN working paper*, 3036094.
- Wang, F. H., Zevin, J., & Mintz, T. H. (2019). Successfully learning non-adjacent dependencies in a continuous artificial language stream. *Cognitive Psychology*, 113, 101223.
- Weyers, I., Männel, C., & Mueller, J.L. (2022). Constraints on infants’ ability to extract non-adjacent dependencies from vowels and consonants. *Developmental Cognitive Neuroscience*, 57, 101149.
- Wilson, B., Spierings, M., Ravignani, A., Mueller, J. L., Mintz, T. H., Wijnen, F., van der Kant, A., Smith, K., & Rey, A. (2020). Non-adjacent dependency learning in humans and other animals. *Topics in Cognitive Science*, 12(3), 843-858.

**Acknowledgements**

This work was supported by the International Centre for Language and Communicative Development (LuCiD) at Lancaster University and the University of Liverpool, funded by the Economic and Social Research Council (UK) [ES/S007113/1].

### Supplementary Information

#### S1. Pre-registered SEM: Separate constructs for vocabulary and grammar at 54 months

As anticipated, the preregistered structural equation model did not provide a good fit to the data, RMSEA = .190, CFI = .784, TLI = .568,  $\chi^2(10) = 31.960$ ,  $p < .001$ . The loadings for the model are shown in Figure S1. Though the relations between the 17-month measures and the grammar latent variable are difficult to interpret, there is a marginal negative relation between the 17-month grammar task and vocabulary at 54 months, indicating that children who look longer at foil over target generalisation sequences tend to have higher vocabulary scores.

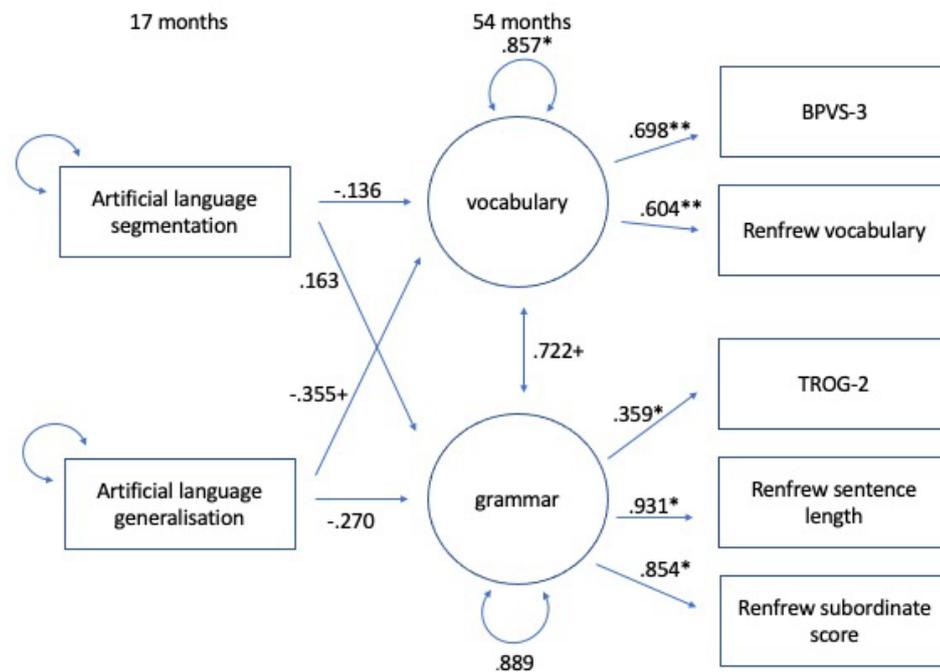


Figure S1. Relations between artificial language segmentation and generalisation at 17-months and natural language skills at 54 months. Values on paths are standardised estimates. +  $p < .1$ , \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ .

## **S2. Additional structural equation model including working memory as a predictor of language skills.**

The model did not provide a good fit to the data, RMSEA = .027, CFI = .842, TLI = .754,  $\chi^2(16) = 31.936, p = .010$ . The loadings for the model are shown in Figure S2. The working memory measure did not relate significantly to the language ability variable. The relation between artificial language generalisation and language ability remained significant, and the relation between artificial language segmentation and language ability remained not significant, consistent with the model in the main paper.

We further conducted correlations between the working memory measure and the artificial language segmentation and generalisation measures. WM correlated positively, but not significantly, with artificial language segmentation,  $r(55) = .24, p = .067$ , and negatively, but again not significantly, with artificial language generalisation  $r(47) = -.25, p = .089$ . Thus, though it is possible that working memory contributed to artificial language learning performance, but did not significantly predict learning, and nor did it mediate the relation between artificial language learning and natural language skills.

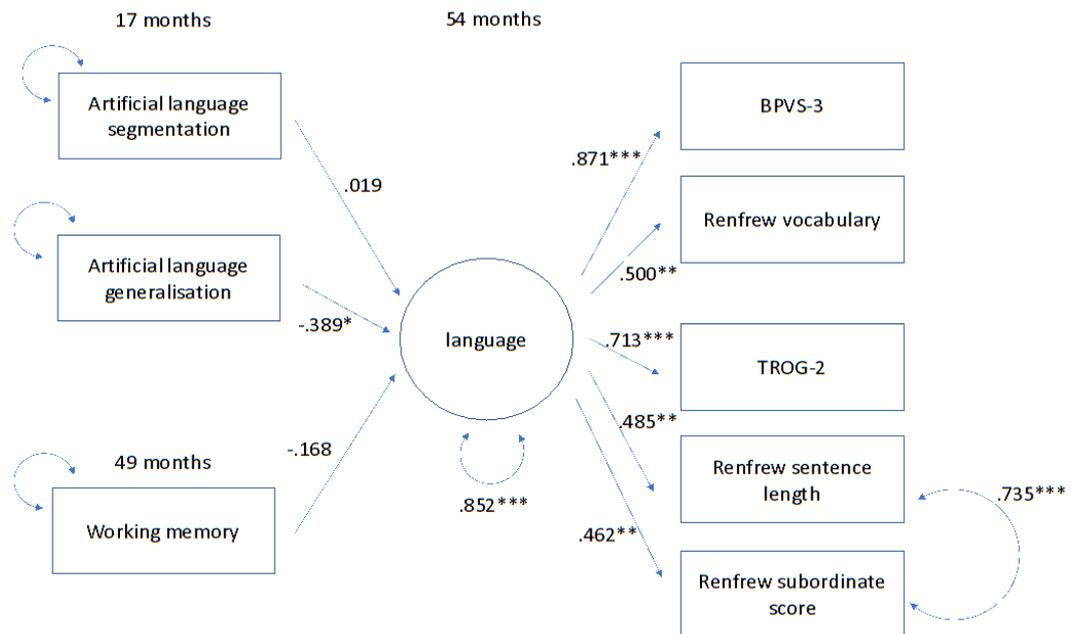


Figure S2. Relations between artificial language segmentation and generalisation at 17-months and natural language skills at 54 months, including also working memory at 49 months. Values on paths are standardised estimates. \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ .