Non-adjacent dependency learning in infancy, and its link to language development

Rebecca L. A. Frost[1*], Andrew Jessop[1], Samantha Durrant[2], Michelle S. Peter[2], Amy Bidgood[3], Julian M. Pine[2], Caroline F. Rowland[1, 2] & Padraic Monaghan[4, 5]

[1]Max Planck Institute for Psycholinguistics, [2]University of Liverpool, [3]University of Salford, [4]University of Amsterdam, [5]Lancaster University.

*Corresponding author:

Rebecca L. A. Frost

Language Development Department

Max Planck Institute for Psycholinguistics

Nijmegen

6525 XD

Netherlands

E-mail: rebecca.frost@mpi.nl

Tel: +31 243521451

Abstract

To acquire language, infants must learn how to identify words and linguistic structure in speech. Statistical learning has been suggested to assist both of these tasks. However, infants' capacity to use statistics to discover words and structure together remains unclear. Further, it is not yet known how infants' statistical learning ability relates to their language development. We trained 17-month-old infants on an artificial language comprising non-adjacent dependencies, and examined their looking times on tasks assessing sensitivity to words and structure using an eye-tracked head-turn-preference paradigm. We measured infants' vocabulary size using a Communicative Development Inventory (CDI) concurrently and at 19, 21, 24, 25, 27, and 30 months to relate performance to language development. Infants could segment the words from speech, demonstrated by a significant difference in looking times to words versus part-words. Infants' segmentation performance was significantly related to their vocabulary size (receptive and expressive) both currently, and over time (receptive until 24 months, expressive until 30 months), but was not related to the rate of vocabulary growth. The data also suggest infants may have developed sensitivity to generalised structure, indicating similar statistical learning mechanisms may contribute to the discovery of words and structure in speech, but this was not related to vocabulary size.

Key Words: language acquisition, artificial grammar learning, speech segmentation, individual differences, statistical learning, vocabulary development

1. Introduction

To reach linguistic proficiency, infants must master two critical tasks; identifying words in speech, and discovering the constraints that shape the way those words are used. Although speech contains no absolute cues to word boundaries (Aslin, Woodward, LaMendola, & Bever, 1996) or grammatical structure (Monaghan, Christiansen, & Chater, 2007), it is replete with distributional information that could assist with these tasks: regular co-occurrence of particular syllables provides a helpful description of what constitutes specific words in a language, whereas information about how words are used in combination helps illustrate how that language operates in terms of its grammatical structure. The ability to draw on such information (*statistical learning*) has therefore been suggested to play a key role in language acquisition (e.g., Conway, Bauernschmidt, Huang, & Pisoni, 2010; Kidd & Arciuli, 2016; Lashley, 1951; Redington & Chater, 1997; Rubenstein, 1973).

Infants have been found to be highly capable of detecting distributional statistics in speech. Indeed, children as young as 8 months old can compute transitional probabilities between syllables (Saffran et al., 1996), and use them to identify word boundaries in a stream of new words (e.g., Aslin, Saffran, & Newport, 1998; Pelucchi, Hay, & Saffran, 2009). At around the same age, infants can also discover simple distributional structure in artificial speech (i.e., an ABA or AAB structure e.g., Marcus, Vijayan, Rao, & Vishton, 1999; Gerken, 2006; 2010), with this capacity potentially increasing in sophistication over development (see e.g., Gómez, 2002; Gómez & Gerken, 1999; Gómez & Maye, 2005; Lany & Gómez, 2008; Lany, Gómez, & Gerken, 2007; Marchetto & Bonatti 2013; 2015, for work on infants' developing ability to track dependencies between adjacent and non-adjacent items).

Taken together, these lines of research provide converging evidence that infants can draw on the statistical properties of language to learn about multiple linguistic features. Further, these studies suggest that learners may develop the capacity to employ statistical learning mechanisms for discovering words and basic structure at relatively similar points in development (though see e.g. Frost & Monaghan, 2016; Peña et al., 2002; and Perruchet, Tyler, Galland, & Peereman, 2004 for the debate concerning the nature of the statistical mechanisms these tasks employ). However, to our knowledge, infants' ability to perform these tasks together during learning remains to be demonstrated (see Marchetto & Bonatti 2013; 2015). In the current study we address this directly, and test whether 17-month-old infants can discover both word boundaries and linguistic structure (non-adjacent dependencies) together, using co-occurrence statistics alone. Further, we examine the way that infants' ability to do so relates to their language development outside of the laboratory.

1.1 Testing acquisition of words and linguistic structure

In both infant and adult research, learners' capacity for joint acquisition of words and language structure has been assessed using artificial languages comprising non-adjacent dependencies – statistically reliable relationships between two items that are separated in speech. Non-adjacent relationships are pervasive in language, and exist at multiple levels of language structure, including syntax (i.e., the relationship between the auxiliary verb and the present participle verb form in the sun is shining), morphosyntax (i.e., co-occurring prefixes and suffixes, e.g., uncovered, independently), and number agreement (i.e., the lion at the zoo roars, the penguins at the zoo swim). Artificial grammars comprising words with morphological non-adjacent dependencies (with an AXC structure, where A and C reliably co-occur, regardless of X) provide the ideal platform for assessing word and structure learning together, since they contain sequences that learners need to discover (words, i.e., AXC strings), as well as structural regularities (i.e., A_C relationships, see e.g., Frost & Monaghan, 2016; Marchetto & Bonatti, 2013, 2015; Peña et al., 2002; and Perruchet et al., 2004, for assessments of word and structure learning using AXC-style input).

To examine the way that word and structure learning proceed in infants, Marchetto and Bonatti (2013; 2015) trained infants on an artificial language with an AXC structure, and examined their capacity for segmenting words from speech, and generalising the internal morphological structure these words contained (the A-C relationships). Learning was tested with the head turn preference paradigm, indexed by differences in looking times to different items at test[1]. In their first set of studies, Marchetto and Bonatti (2013) examined whether the emergence of morphosyntactic structure learning actually precedes statistical segmentation, with the view that the former may act as an economical solution to identifying possible word candidates in speech. To this end, they pitted adjacent transitional probabilities against nonadjacent dependency structure at test, and examined whether infants relied on one type of information over the other to identify likely word candidates in speech. They report that 12- and 18-month-olds preferentially drew on within-word structure (A-C relationships) when speech was segmented with pauses. When speech was continuous (as is more typical of natural language), 18-month-olds relied more on transitional probabilities to identify likely words, whereas 12-month-olds showed no preference (they did not discriminate between the two types of test item).

Their second set of studies explored related issues in 7- and 12-month-olds, and found that at 12 months, infants could use statistical relationships between syllables to extract words from continuous speech, but could only detect the non-adjacent dependencies contained in the words if speech was segmented (Marchetto & Bonatti, 2015). Seven-month-olds were unsuccessful at both segmenting words and generalising A-C structure, though they were able to discriminate between words and non-words after exposure to a segmented version of the speech stream. The authors concluded that infants' capacity for learning can be seen to critically develop over time, such that by 12 months infants possess the cognitive resources to

[1] In Marchetto and Bonatti's (2013) study with 12- and 18-month olds, sensitivity to words and structure was assessed together with part-word versus rule-word comparisons; part-words occurred in speech with relatively higher frequencies than other items such that they were sound candidates for lexical items, and rule-words comprised an A-C dependency with an intervening A or C from another pairing, such that they were grammatical, but new. Inferences about infants' performance were based on the direction of infants' looking preferences for these comparisons. While infants did attend differentially to these stimuli, unpacking the nature of this difference is difficult (see related comments about interpreting looking preferences in the Results and Discussion sections), particularly given the combined assessment of these skills. In their 2015 study with 7- and 12- month olds, sensitivity to words and structure was examined separately using looking times to words versus non-words (to assess segmentation) and rule-words versus non-words (to assess structure learning).

analyse the internal composition of words (but can only do so if the speech contains information that aids segmentation).

These studies provide an informative first foray into infants' capacity to discover words and structure using the statistical information contained in speech. Yet, in each of these studies, a small number of design features make performance somewhat difficult to unpack (see footnote 1 for an overview of the test stimuli). Further, due to methodological differences across the sets of studies, it is difficult to compare these datasets - meaning the developmental trajectory for these tasks is yet to be conclusively established. Thus, further research is needed to understand infants' ability to discern statistically defined words and structure from speech – both in terms of the nature of these processes, and when and how they develop.

Nevertheless, when viewed together these data suggest that the discovery of statistically-defined words and (word-internal) structure may be underpinned by different processes (i.e., statistical segmentation versus algebraic computation of structure), each with a slightly different developmental trajectory. This proposed distinction in processing between word- and structure- learning is consistent with much of the adult literature on this topic (e.g., Peña et al., 2002), and is in line with the suggestion that learners perform these tasks separately during language acquisition, drawing on separable and distinct computations for segmenting speech versus generalising structure (Marcus et al., 1999, Peña et al., 2002). However, recent advances in the adult literature have highlighted a possible methodological confound which may have influenced performance in the research that generated these conclusions. Specifically, in prior studies, generalisation of structure was typically assessed with comparisons involving "rule words" - a familiarised A-C dependency, with an intervening A or C element from a different dependency – versus an item that is infrequent or absent from the training speech. Structural generalisation would be evidenced by a preference for rule words over the competitor item on a 2AFC test with adults, or a difference in looking times to rule words and part-words/non-words in infant head-turn preference studies.

Though such comparisons permit assessment of preference for the overall structure, they require learners to use trained A and C items flexibly in a way that conflicts with their knowledge of where those syllables should occur within sequences. Frost and Monaghan (2016) argued that this may have constrained learners' willingness to generalise, and suggested that learners may be able to do so in the absence of such conflict. Indeed, using amended generalisation stimuli (containing entirely novel intervening items, rather than repositioned A or C items), Frost and Monaghan (2016) demonstrated that adults could learn about words and linguistic structure at the same time, in the absence of additional information such as pauses between words (see Frost, Isbilen, Christiansen, & Monaghan (2019), and Isbilen, Frost, Monaghan, and Christiansen (2018) for replications of this effect). Thus, the processes underlying word and structure learning may be more similar in nature than previously suggested, with statistical learning about words and linguistic structure possibly being served by the same (or at least the same type of) mechanism.

With this in mind, it is possible that infants' true capacity for learning about non-adjacent dependencies from continuous speech may not have been detected in previous studies, perhaps due to limitations of the learning measure, rather than the learner. We propose that implementing methodological changes to the stimuli in line with those made by Frost and Monaghan (2016) would provide a closer approximation of infants' capacity to generalise non-adjacent dependencies, shedding further light on the developmental trajectory for these tasks.

1.2 Statistical learning, language development, and individual differences

A key question in interpreting data from artificial language learning studies is what performance on these tasks actually means in terms of natural language development. That is, how does infants' ability to detect patterns in an artificial grammar relate to how they learn language in the world outside of the laboratory? Recent research with adults has indicated that participants' ability to compute statistics over artificial grammars relates to their competence on other linguistic tasks, shedding light on the way statistical learning skills may shape or reflect language learning more broadly. For instance, Isbilen et al. (2018) demonstrated that adults' capacity to compute statistically defined non-adjacent dependencies relates to their ability to learn more naturalistic language structure on a cross-situational learning task that taught learners a small-world version of Japanese.

Emerging evidence for the role of statistical learning in language acquisition also comes from literature on individual differences, which seeks to determine whether variation in learners' performance on experimental language learning tasks relates to variation in natural language skills. There is growing support for the existence of a meaningful relationship between an individual's statistical learning ability and their "real-world" language skills for both children (e.g., Kidd, 2012; Kidd & Arciuli, 2016) and adults (e.g., Christiansen, Conway, & Onnis, 2012; Conway et al., 2010; Misyak, Christiansen, & Tomblin, 2010), strengthening the possibility that statistical computations play a role in natural language acquisition.

Recent work by Lany (2014) and Lany and Shoaib (2019) shed new light on this relationship by demonstrating that infants' performance on a statistical language learning task differed as a function of their natural language ability. Lany (2014) tested infants' ability to map distributional information onto semantic categories (animals and vehicles), then examined whether their capacity to do so was related to their vocabulary size. Co-occurrence of determiners and nouns during familiarisation was found to inform infants' formation of semantic categories, helping them to use the new nouns as labels. However, this was only the case for infants with higher scores on the MacArthur-Bates CDI measure of grammar development (Fenson et al., 2007) – providing a promising indication that infants' capacity for statistical learning relates to their language learning outside of the lab, with more advanced users of natural language outperforming their peers on the statistical learning task.

Further, Lany and Shoaib (2019) found evidence to suggest that for some 15-month-olds, their ability to learn non-adjacent dependencies in an artificial language learning task (dependencies between words in segmented speech, e.g., Gómez, 2002) may be related to their vocabulary size at the time of testing, and possibly at prior and subsequent points in development (at 12 and 18 months). For some participants, there was also evidence that performance at 15 months predicted later sensitivity to analogous non-adjacent dependencies in natural language (tested at 18 months). However, the effects in this study were complex, with substantial differences across sexes, and the relationships described here were not observed uniformly across participants - most correlations were only observed for the small sub-sample of females (Nrange = 10-16).

Consequently, more research with different statistical structures, and different age groups, is needed to understand the relationship between statistical learning and language development more fully. Here, we contribute to this literature by examining whether infants' capacity for statistical segmentation relates to their vocabulary size.

An important step in understanding the role of statistical learning in language acquisition is to look at how it relates to other language skills across time, over development, as well as concurrently. There is a growing body of literature suggesting that infants' early linguistic skills relate to their subsequent language development (typically indexed by CDI scores) – giving critical insight into the extent to which particular linguistic skills serve language learning more broadly. For instance, research on phonetic perception suggests that infants' behavioural (Tsao, Liu, & Kuhl, 2004) and neural responses (Molfese, 2000; Molfese & Molfese, 1985, 1997; Rivera-Gaxiola, Klarman, Garcia-Sierra, & Kuhl, 2005) to phonemic speech sounds may play a role in explaining the language skills of those children at later points in development (see Cristia, Seidl, Junge, Soderstrom, & Hagoort, 2014 for a review).

Similarly, research on speech segmentation has found that infants' recognition of new words in spoken utterances relates to their vocabulary development, and this relationship has been shown for both behavioural (Newman, Bernstein Ratner, Jusczyk, Jusczyk, & Dow, 2006; Newman, Rowe, & Bernstein Ratner, 2016; Singh, Reznick, & Xuwhua, 2012) and neural (Kidd, Junge, Morrison, Spokes & Cutler, 2018; Junge, Kooijman, Hagoort, & Cutler, 2012) indices of speech segmentation. Research has also found a relationship between laboratory-based word learning ability and vocabulary size (Bion, Borovsky, & Fernald, 2013).

For statistical learning, though, it is not yet known how infants' performance on laboratory tests of word segmentation and structure learning relates to their language development (but for related preliminary evidence, see Lany and Shoaib’s (2019) study of nonadjacency learning from pre-segmented speech). While much research has documented infants’ ability to draw on statistics in speech to learn about words and within-word structure, infants have never been found to be capable of performing both tasks together from statistics alone. Further, it is not yet known how infants’ performance on these tasks relates to different aspects of language development. For instance, word learning and structure learning may be separable processes (Marchetto & Bonatti, 2015), in which case we might expect only statistical segmentation to relate to vocabulary development (that is, if it relates at all to natural language learning). Alternatively, if word learning and structure generalisation involve related processes, both might relate to vocabulary development.

We thus extended the work of Marchetto and Bonatti (2013; 2015) and Frost and Monaghan (2016), to examine whether infants, like adults, can compute word-like and structure-like regularities at the same time, from the same set of distributional statistics - without any extra cues in the speech signal (i.e., pauses between words). Demonstrating that infants are able to detect both words and structure from this input would have important implications for the simultaneity of these tasks in language acquisition, and the statistical computations that may underlie them.

Importantly, we also tested whether infants' statistical learning ability related to their natural language ability, both concurrently and over development; if the transitional information contained within speech does support natural language acquisition (or if children's language development supports their ability to compute over the statistical information contained in speech), it follows that infants' capacity for statistical learning on this task may be related to their language development in the world outside of the laboratory. This relationship could take two forms; statistical learning ability may relate to vocabulary size (i.e., children with a greater capacity for statistical learning may have larger vocabularies), and it may also relate to vocabulary growth (i.e., children with greater statistical learning ability may increase their vocabulary more rapidly over development). To test this possibility, we examined whether performance related to a measure of natural language development taken at the time of testing (UK-CDI, Alcock, Meints, & Rowland, 2020), and at 19, 21, 24, 25, 27, and 30 months (Lincoln CDI, Meints & Fletcher, 2001).[2]

We expected that infants would be able to segment the speech, and generalise the language structure to novel consistent items (Frost & Monaghan, 2016). Further, we expected that infants' performance on the segmentation task would relate to their concurrent vocabulary scores (Junge et al., 2012; Kidd et al., 2018; Lany, 2014; Newman et al., 2006; Newman et al., 2016; Singh et al., 2012), and possibly vocabulary growth over time. Testing whether generalisation of the artificial language also relates to vocabulary development provides insight into the similarities or potential distinctions between word learning and grammatical generalisation.

[2] Over the Language 0–5 Project, measures of vocabulary size were taken at a range of time points. Between 8 and 18 months, caregivers completed the UK-CDI Words and Gestures (UK-CDI; Alcock et al., 2020). The UK-CDI is suitable for use up until 18 months, so for subsequent time-points (between 19 - 30 months) caregivers completed the Lincoln CDI Words and Sentences (Meints & Fletcher, 2001).

2. Method

2.1 Participants

The experiment was completed by 71 infants (40 females, 31 males; aged between 16.5 – 17.5 months, mean age = 517 days), recruited from Liverpool, UK. All infants were monolingual native English learners, born at term, with normal vision and hearing. All infants were typically-developing at the time of testing. Infants were tested in the laboratory at The University of Liverpool.

This study forms part of a larger longitudinal project in the North West of England, the Language 0-5 Project (Rowland, Bidgood, Durrant, Peter, Pine, unpub). Ninety-five families were recruited to take part. Of these, one family was excluded due to responses on a family background questionnaire (persistent ear infections likely to affect hearing) and four withdrew before the project began. This resulted in a final sample of 90 families. Out of the final 90 families, nine had a family history of language delay or dyslexia. More general information about sample background can be found in Peter et al., (2019).

2.2 Design

The data were collected as part of a large-scale study of language development and individual differences in language acquisition. Due to the unique requirements of group-level and individual differences level research, studies attempting to assess both must typically prioritise one over the other. Here, we prioritised assessment of individual differences, and controlled for this statistically to test for effects at the group level. Thus, to minimise task-related variance across learners, all participants received the same stimuli, in the same order (see Procedure section for further information regarding this decision, and see the results section for details of how we controlled for this in our analysis). Ethical approval was given by the University of Liverpool Research Ethics Subcommittee for Non-Invasive Procedures (RETH000764) for the project.

2.3 Materials

2.3.1 *Stimuli*

Speech stimuli were created using Festival speech synthesiser (Black, Taylor, & Caley, 1990) and were based on those used by Frost and Monaghan (2016; see also Marchetto & Bonatti, 2013; 2015, and Peña et al., 2002). The language contained six monosyllabic items (*ba (*/bɑ/*), mu* (/mu/), *so* (/saʊ/)*, li* (/li/), *ga* (/gɑ/), *fe* (/feɪ/) which were used to create two non-adjacent pairings [ba-so] and [li-fe] with two possible X items [mu] and [ga] which intervened the dependencies ($A_1X_{1-2}C_1$ and $A_2X_{1-2}C_2$), as in Marchetto and Bonatti (2013; 2015). Phonemes used for A, X and C items contained a mix of plosives and continuants, since similarities in phonological properties of non-adjacent dependent syllables have been shown to support acquisition of those dependencies (Newport & Aslin, 2004; but see Frost et al., 2019, and Onnis, Monaghan, Richmond, & Chater, 2004 for evidence this is not essential for learning). Each AXC string lasted approximately 700ms. There were four additional syllables containing a mix of plosive and continuant consonants which were reserved for testing generalisation to novel items (*ni* (/ni/)*, po* (/paʊ/), *du* (/du/)*, ve* (/veɪ/)).

2.3.2 Training

A 15-minute-long continuous speech stream was created using the Festival speech synthesiser (Black et al., 1990) by concatenating the four AXC words (*bamuso*, *bagaso*, *limufe*, *ligafe*)[3]. This was produced using a female voice at 140 hz, with the constraint that no $A_iX_jC_i$ sequence was immediately repeated. In the speech stream, transitional probabilities for A-C syllables were always 1, while probabilities for A-X and X-C transitions were .5. The likelihood that a particular AxC word would be followed by another given word was .33. The

[3] Marchetto and Bonatti (2013; 2015) used comparatively shorter training streams (~2-3 min), in keeping with the familiarisation paradigm they employed during exposure. Since the study at hand trained participants via incidental learning (to minimise the amount of time spent at the eye tracker prior to the test, in an effort to optimise inclusion rates) we used a longer training stream, in line with the standard procedure for this type of exposure (see e.g., Saffran, Newport, Aslin, Tunick, & Barrueco, 1997). It is possible that this may lead to shorter overall looking times than those observed in Marchetto and Bonatti's (2013; 2015) studies, but this should not compromise our ability to compare looking across the two types of test-items on each task.

speech stream was edited to have a 5 second fade in and out, so that the onset and offset of the speech could not be used as a cue for segmentation.[4]

2.3.3 Testing

We assessed segmentation by measuring looking times to two types of trials; *words* and *part-words. Word* trials comprised repetitions of one of the words used in the familiarisation stream (e.g., *bamuso bamuso bamuso...*). *Part-word* trials contained repetitions of items that occurred in the training speech but straddled word boundaries, comprising the last syllable of one word and the first two syllables of another word (C_iA_jX; *sobamu, feliga*), or the last two syllables of one word and the first syllable of another (XC_iA; *gasoli, mufeba*). There were therefore four word trials and four part-word trials, each of which was presented twice, giving 16 trials in total (see Marchetto & Bonatti, 2015).

We assessed generalisation with trials containing repetitions of *rule-words* and *non-words* (Frost & Monaghan, 2016; Marchetto & Bonatti, 2015). Rule-words comprised an $A_i\_C_i$ non-adjacency, intervened by one of the four novel syllables (so, taking the form A_iNC_i, where N indicates the novel syllable; *baniso*, *baposo*, *lidufe*, *livefe*). Non-words were part-words in which one syllable was replaced with a novel syllable, to ensure any preference observed on generalisation trials could not be attributed to the presence of a novel syllable alone. Novel syllables could appear in the initial or final position (so, taking the form C_iA_jN; *solive*, *febadu*, or NC_iA_j; *posoba*, *nifeli*) with two trials adhering to each possible non-word structure. There were therefore four trials of each item type, and each was repeated twice, giving 16 trials in total (see Marchetto & Bonatti, 2015).

The presentation order of trials was pseudo-randomised using the same criteria as in Marchetto and Bonatti's study (2015), with no immediate repetition of particular items, and a

[4] In line with Marchetto & Bonatti (2013), and to reduce task-related variance (which may compromise assessment of individual differences), we used one artificial language. See Marchetto and Bonatti (2015) and Frost and Monaghan (2016) for evidence that using multiple counterbalanced input streams would not have impacted group-level results.

maximum of three consecutive trials of the same type (with regard to both word type, and left/right location of stimulus presentation). The precise presentation order for each task is given in the supplemental materials.

2.4 Procedure

Infants were familiarised with the experimental language for 15 minutes via incidental learning (Gómez, Bootzin, & Nadel, 2006; Saffran, Newport, Aslin, Tunick & Barrueco, 1997), with infants playing quietly with the experimenter (i.e., with no verbal communication) while the speech stream played at a comfortable volume in the background. During the incidental learning phase, caregivers completed questionnaires for another component of the Language 0-5 project (these questionnaires are not relevant for the current study, so will not be discussed further).

Following familiarisation, we assessed infants' learning using an adaptation of the classic head turn preference paradigm (Kemler Nelson et al., 1995), modified to incorporate an eye-tracker, which measured infants' looking times to each test trial. Eye movements were recorded using an SR Research Eyelink 1000 plus (SR Research: Ottawa, Ontario, Canada) in remote mode using the remote arm configuration with a target sticker, which permits stable tracking while accommodating some level of movement. Infants were seated in a car seat in front of the eye-tracker (affixed to a 17" LCD monitor), which was uniquely positioned for each child such that the display distance was 580-620mm. Trials began after successful five-point calibration.

Sound stimuli were played through speakers positioned behind the monitor, to the left and right sides of the screen. Test items were paired with a visual stimulus (an animated clip of a slow-moving hand, as in Marchetto & Bonatti, 2015), set against a black background, which appeared onscreen on either the left or the right, in accordance with the location of the sound. Individual test trials occurred twice; once to the left, and once to the right. On each trial, infants heard repetitions of a test-item, separated by a 500ms pause, with items played in the same voice and at the same rate as in familiarisation. Trials could last for a maximum of 65 seconds (Marchetto & Bonatti, 2013, 2015), and were gaze contingent, such that trials terminated if an infant looked away from the visual stimulus for more than 2 seconds. After each trial ended, a fixation stimulus appeared at the centre of the screen to re-direct infants' attention, and the next trial began after infants had attended to this for 2 seconds.

To minimise item and task order related variance, which would add noise to our individual differences analyses, all infants completed the segmentation trials first followed by the generalisation trials, and all trials were presented in the same pseudo-randomised order (for more justification of this decision, which is necessary for adapting group-based experimental procedures for individual differences designs, see e.g. Panter et al., 1992 for a summary of the effects of item and test order on the reliability of comparisons across individuals, and see Cooper et al., 2017, for details of how to apply these considerations to individual differences research in cognition. For information on how we controlled for this statistically in our group-level analysis, see the Results section). The segmentation and generalisation phases were separated by a brief comfort break, during which infants watched a short cartoon (a 135 s excerpt of Pingu - chosen for its lack of linguistic content). For each infant, familiarisation and testing took place in the same laboratory. Caregivers were asked to refrain from communicating verbally with their infant during both familiarisation and testing, and were asked to avoid directing infants' attention at test.

This experiment formed part of a large longitudinal cohort study assessing language development in children. Thus, infants tested for this study had participated in studies assessing various aspects of language learning prior to this session. However, none of these studies contained the same words or grammar-like rules as this study, and none of them examined infants' capacity for statistical language learning. On the day of testing, infants did not complete any other behavioural studies prior to participating in the study at hand.

3. Results

3.1 Data preparation

Filtering criteria were applied to the data: Trials shorter than 700 ms (the approximate length of a test item) were excluded from analysis, as were trials with looking times greater than 2SD beyond the mean looking time for that trial. In their study, Marchetto and Bonatti (2015) excluded trials with looking times shorter than 1000 ms. However, we implemented a lower minimum cut-off to maximise the amount of useable trials, and to align this cut-off with the stimuli such that looking times could be more confidently linked to attending to the test items. All data that permitted comparison of looking to the different types of experimental trials were included in the analysis; that is, infants were only excluded if they failed to provide data for at least one of each trial type after the data were filtered. For *segmentation*, data for 70 participants was included in the analysis, and the mean number of trials included per child = 11 (range = 2-16). For *generalisation*, data for 61 participants was included in the analysis, and the mean number of trials included per child = 9 (range = 2-14). See the supplemental materials for replications of the main group-level analyses for each task with the full, unfiltered datasets (all critical effects are replicated with the raw dataset).

3.2 Data Analysis

We first examined infants' performance on the segmentation and generalisation trials, assessing looking behaviour on each of these tasks separately. We then examined whether infants' performance on these tasks related to their concurrent CDI scores. In subsequent analyses, we investigated the relationship between statistical learning ability (speech segmentation) and vocabulary development over time.

Note that for both segmentation and generalisation, due to possible effects of trial order we do not make inferences based on overall group means; instead, we report pre-planned analyses that control for trial order statistically (see sections 3.2.1, 3.2.2, and 3.2.3).

3.2.1 Segmentation

Overall, infants' average looking time for *word* trials was $M = 4346.33$ ms (SD = 3663.24), and for *part-word* trials was $M = 3498.33$ ms (SD = 2744.31).

Linear mixed-effects analysis was performed on the data for the segmentation trials (Baayen et al., 2008), which modelled the probability (log odds) of looking times considering variation across participants and materials, as well as across the two types of test items (words and part-words), to determine whether these differentially affected looking behaviour.

The model was built incrementally, and was initially fitted specifying random effects of subject, gender, and stimuli location, to account for variation in performance across participants and across items displayed on either the left or right of the screen. Random intercepts and slopes were omitted if the model failed to converge with their inclusion. We then added fixed effects and interactions for trial order and test item type; these were added incrementally, and were retained in the model if significant. Trial order was included as a fixed effect as we predicted a habituation-related decline in looking times to stimuli over the course of the task (it was also important to control for this given that all infants received the same trial order, due to the individual differences nature of the design). Importantly, we added this to our model first so that subsequent comparisons could test whether there was a difference in looking to word versus part-word trials *over and above* any effects of trial order.[5] Experimental effects are thus effects that are observed once variation associated with order has been accounted for, and are therefore not due to performance on any particular trial. A summary of the final model (i.e., the most complex model at the end of this incremental process) is reported in Table 1.

The linear mixed-effects analysis revealed a significant effect of trial order, with looking times decreasing as anticipated over the course of the session (model fit improvement over model containing random effects: $\chi^2(1) = 105.48$, $p < .001$). Crucially, there was a

[5] For supplementary exploratory analysis which statistically controls for trial order differently, through residualisation, see the supplementary materials. All effects were replicated using this approach.

significant effect of trial type, over and above the effect of trial order, indicating that infants responded differently to words and part-words (model fit improvement over model containing random effects and a main effect of trial order: χ^2 (1) = 5.128, p = .023), suggesting that they had segmented the words from the speech stream.

Likelihood ratio test comparisons indicated that model fit was significantly improved when we added the interaction term for trial type and trial order, with infants' looking times to words and part-words changing over the course of the task (model fit improvement over model containing just main and random effects: χ^2 (1) = 9.843, p = .002). This interaction is likely to be a product of habituation, as the difference in looking times for word versus part-word trials reduces over the course of the session (see Figure 1).

Table 1. Summary of the linear mixed-effects model of (log odds) looking times on the segmentation trials.

Fixed effects	Estimated coefficient	SE	Wald confidence intervals 2.50%	97.50%	t value
(Intercept)	4824.26	405.97	4028.576	5619.948	11.883
Trial	-158.44	33.96	-224.995	-91.883	-4.666
Trial type	1631.96	424.41	800.142	2463.785	3.845
Trial * Trial type	-146.72	47.04	-238.925	-54.515	-3.119
Random effects	Variance	Std. Dev.			
Subject (Intercept)	750093	866.1			
Trial_type (slope)	47847	218.7			
L_or_R (Intercept)	93709	306.1			

784 observations, 70 participants. R syntax for the final model is: lmer (total_looking ~ trial*seg_trial_type + (1|L_or_R) + (1+seg_trial_type|subject), data = seg_filt_data, REML = TRUE)

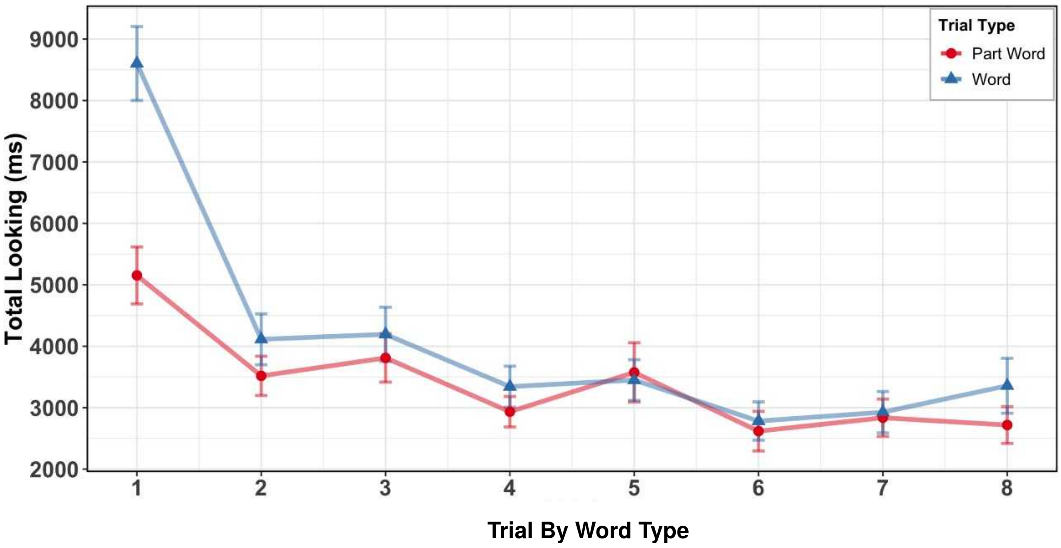

Figure 1. Mean overall looking times to word and part-word trials over the task, with SE. Trial number (1-16) is broken down into trial number by trial type (1-8), to illustrate the relative difference between looking on the first, second, *etc.* trial of each type (though we did not alternate perfectly between types of trial, and trial order was statistically controlled for in the analysis). We note that individual differences analysis (reported in section 3.2.4) shows that segmentation performance was not homogeneous; see this section for a visualisation of looking behaviour split by looking preference, and for evidence of the stability of the effects over the task (thus, initial trial performance does not drive the observed effects of trial type). See supplementary figure iii for an illustration of looking behaviour residualised against trial.

3.2.2 Generalisation

Overall, infants' average looking time for *rule-word* trials was *M* = 3030.70 ms (SD = 2374.70), and for *non-word* trials was *M* = 3663.46 ms (SD = 3025.75).

Linear mixed-effects analysis was performed on the data for the generalisation trials (Baayen et al., 2008), which modelled the probability (log odds) of looking times considering variation across participants and materials, and across the two types of test items (rule-words and non-words), to determine whether infants looked differently to rule-words and non-words at test. A summary of the final model is reported in Table 2. As with the segmentation analysis, the model was built incrementally, and was initially fitted specifying random effects of subject, gender, and stimulus location, to account for variation in performance across participants and across items displayed on either the left or right of the screen. Random effects and slopes were omitted if the model failed to converge with their inclusion. We then added fixed effects and interactions for trial and test item type, with significant main effects/interactions being retained in the model.

As expected, there was again a significant effect of trial order, with looking times decreasing over the course of the session (model fit improvement over model containing random effects: χ^2 (1) = 24.011, p < .001). Critically, there was a significant effect of trial type indicating that infants responded differently to rule-words and non-words (model fit improvement over model containing random effects and a main effect of trial: χ^2 (1) = 8.626, p = .003), suggesting that infants were sensitive to the structure of the words in the speech stream (see Figure 2).

Again, likelihood ratio test comparisons indicated that model fit was significantly improved when we added trial type, trial order and the interaction term for trial type and trial, with infants' looking to rule-words and non-words changing over the course of the task (model fit improvement over model containing just main and random effects: χ^2 (1) = 6.1982, p = .013). This interaction could relate to a preference switch during the task, as the difference in looking times between trial types seems to fluctuate over the course of the session. Alternatively, this could be due to participants not converging on a stable representation.

In sum, as a whole our sample discriminated between words and part-words in our segmentation task, and between non-words and rule-words in our generalisation task.

Table 2. Summary of the linear mixed-effects model of (log odds) looking times on the generalisation trials.

Fixed effects	Estimated coefficient	SE	Wald confidence intervals 2.50%	97.50%	t value
(Intercept)	4989.306	299.051	4403.176	5575.435	16.684
Trial	-180.353	32.765	-244.571	-116.133	-5.504
Trial type	-1572.388	425.741	-2406.824	-737.951	-3.693
Trial * Trial type	125.89	49.86	28.164	223.609	2.525
Random effects	**Variance**	**Std. Dev.**			
Subject (Intercept)	445891	667.8			
Gen_trial_type (slope)	47285	217.5			

549 observations, 61 participants. R syntax for the final model is: lmer (total_looking ~ trial*gen_trial_type + (1+gen_trial_type|subject), data = gen_filt_data , REML = TRUE)

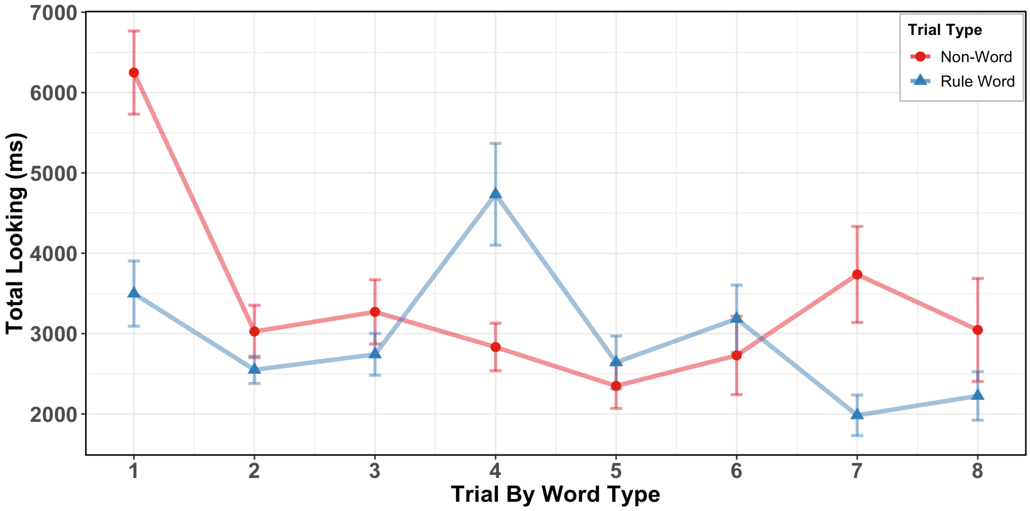

Figure 2. Mean looking times to non-word versus rule-word trials over the course of the task, with standard error. Trial number (1-16) is broken down into trial number by trial type (1-8), to illustrate the relative difference between looking on the first, second, *etc.* trial of each type.

3.2.3. Indexing performance with Cohen's d

In order to examine the relationship between infants' statistical learning skills and their natural language abilities, we required a measure of individual infants' performance on the task, indicating the size of the difference in looking times between test stimuli, but also taking into account the variance in looking times for each child. For this purpose, we computed a measure of effect size (Cohen's d) for each participant. For the segmentation data, these were calculated by subtracting looking times to words from looking times to part-words, then dividing this by the pooled standard deviation of looking times across all segmentation trials (per infant). A positive effect size would indicate a preference for part-words (novelty preference) whereas a negative effect size would indicate a preference for words (familiarity preference). An effect size around zero would indicate no clear preference.

For the generalisation data, effect sizes were calculated by subtracting looking times to rule-words from looking times to non-words, then dividing this by the pooled standard deviation of looking times across all generalisation trials (per infant). A positive effect size would indicate a preference for looking toward the non-words (i.e., a novelty preference to sequences not conforming to the A_C non-adjacency structure), whereas a negative effect size would indicate a preference for looking toward the generalised, rule-word items (thus, a familiarity preference to the A_C structures).

As yet, there is no established method for mapping individual differences in looking preferences onto representations of knowledge. However, Cohen's d allowed us to ascertain differences between individuals, while taking into account individual variation in looking behaviour across the task. Both the size and direction of this difference in looking could be indicative of the nature and extent of learning. For instance, children may demonstrate familiarity or novelty preferences according to the extent to which the stimuli are treated as linguistically relevant, or novel. Note that processing mechanisms that result in patterns of familiarity or novelty preference will exert opposite effects on learning, so observations of no preference in some children may be a consequence of these opposing forces balancing out over the task, instead of tipping the scales in a particular direction. Equally, the size of the effect could indicate learning, with higher scores possibly denoting greater (and consistent) preferences. Here, we assume that children who show a consistent novelty preference across trials have encoded the information better than those who show no overall preference or a consistent familiarity preference, on the basis that a novelty preference indicates better encoding of the familiarised stimuli than no preference or a familiarity preference.[6]

Figures 3 and 4 below illustrate infants' performance on the segmentation and generalisation tasks, respectively, for children with each type of preference (and for children with no overall preference). There was no significant correlation between Cohen's d scores for segmentation and generalisation performance (Pearson's r = -.02, N = 61, *p* = .892),

[6] See Lany & Shoaib (2019) for related individual differences analysis of looking times data with similar assumptions, but using mean difference scores, rather than Cohen's d.

suggesting that infants' looking behaviour on the segmentation and generalisation trials was not statistically related.

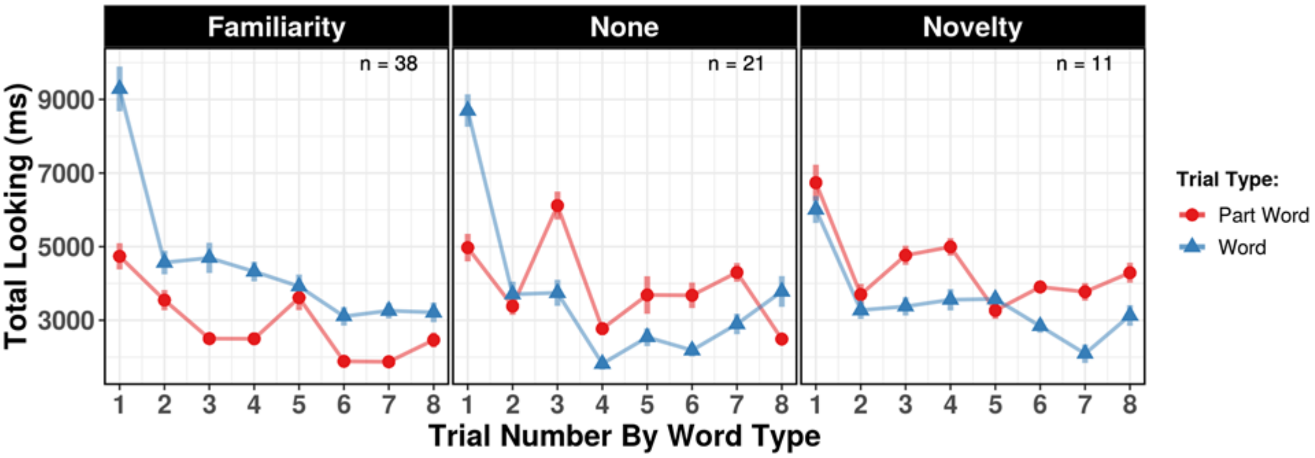

Figure 3. Mean overall looking times to word and part-word trials over the course of the segmentation task with SE, given for participants with a familiarity preference (preferring words, d < -0.2), no preference (d = -0.2 - 0.2), and a novelty preference (preferring part-words, d > 0.2), respectively. Effect size boundaries for defining preference groups were determined based on Cohen (1992). We note the stability of the familiarity and the novelty effects across the task.

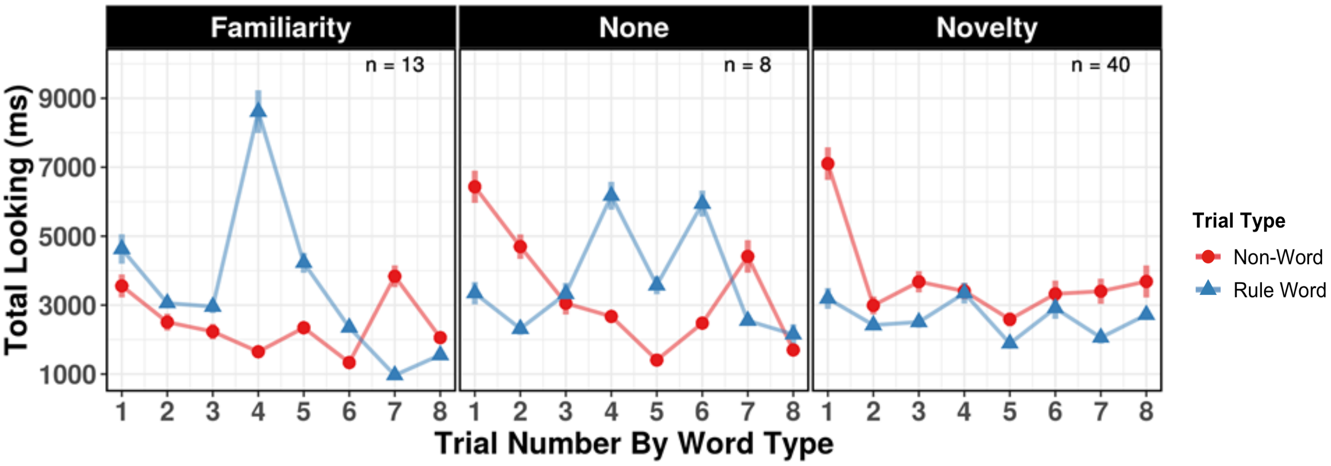

Figure 4. Mean overall looking times to non-word and rule-word trials over the generalisation task with SE, given for participants with a familiarity preference (preferring rule-words, d < -

0.2), no preference (d = -0.2 – 0.2), and a novelty preference (preferring non-words, d > 0.2), respectively. Effect size boundaries were determined based on Cohen (1992).

3.2.4 Statistical learning and vocabulary development

3.2.4.1 Relationship with concurrent vocabulary

We first determined the relationship between infants' statistical learning and their concurrent vocabulary, by correlating infants' statistical learning scores (Cohen's d) with their UK CDI scores (expressive and receptive) at 17 months. CDIs were completed either on the day of the experiment, or within the week prior to testing.

For speech segmentation, children's effects were expressed along a continuum from preferences for familiar words (negative effect) to preferences for novel, part-word stimuli (positive effect), see panels A and B of Figure 5. Overall, there was a significant positive correlation between statistical learning performance and vocabulary size, both for expressive (Pearson's r = .32, N = 68, *p* = .008) and receptive (Pearson's r = .32, N = 68, *p* = .007) scores – suggesting statistical language learning skills and vocabulary may be critically related. Of particular note is the direction of this relationship; data indicate that participants with larger vocabularies (indexed by larger CDI scores) showed larger preferences for novel, rather than familiar, items at test, whereas the opposite was true for children with smaller vocabularies. This result in line with Hunter and Ames' (1988) model, which suggests that children with larger vocabularies are more advanced in their linguistic development than children with smaller vocabularies, and are thus more inclined to show a novelty preference.

To verify the possible maturational distinction between familiarity versus novelty seekers on our segmentation task a ternary split was applied to the data, dividing the sample into familiarity seekers, novelty seekers, and infants with no preference. A Cohen's d of 0.2 is traditionally considered a small effect size (Cohen, 1992). Therefore, children with a Cohen's d of -0.2 or below were classed as having a familiarity preference, while children with a Cohen's d greater than -0.2 but less than 0.2 were classed as having no preference, and children with a Cohen's d of 0.2 or above were classed as having a novelty preference.

We modelled a series of exploratory comparisons on these data to determine whether the direction of a child's preference (or lack thereof) related to their vocabulary size. Bootstrapped (r = 10000) multiple regression models were fit to the concurrent receptive (F(2, 76) = 4.75 [-4.69, 11.42], p = .011, R^2 = 0.11) and expressive (F(2, 76) = 3.19 [-6.61, 10.11], p = .047, R^2 = 0.08) vocabulary scores (square brackets contain 95% CI). These models indicated that children with a familiarity preference had a significantly smaller receptive vocabulary (β = -78.53 [-137.52, -19.39], SE = 30.13, t = -2.61, p = .009) than children with no preference, whereas children with a novelty preference had a significantly larger receptive vocabulary (β = 111.62 [25.74, 197.23], SE = 43.75, t = 2.55, p = .011) than those who had no preference at test (and by extension, than the children with a familiarity preference). For expressive vocabulary, these differences were in the same direction (familiarity preference versus no preference: β = -32.36 [-70.16, 5.55], SE = 19.31, t = -1.68, p = .094; novelty preference versus no preference (β = 51.3 [-18.21, 119.79], SE = 35.21, t = 1.46, p = .145), however these did not reach statistical significance.

For structure generalisation, there was again a continuum of effects, but statistical learning performance and concurrent vocabulary size were not significantly correlated for either expressive (Pearson's r = -.04, N = 59, p = .762) or receptive (Pearson's r = -.09, N = 59, p = .508) scores (see Figure 5, panels C and D), and were not explored further.

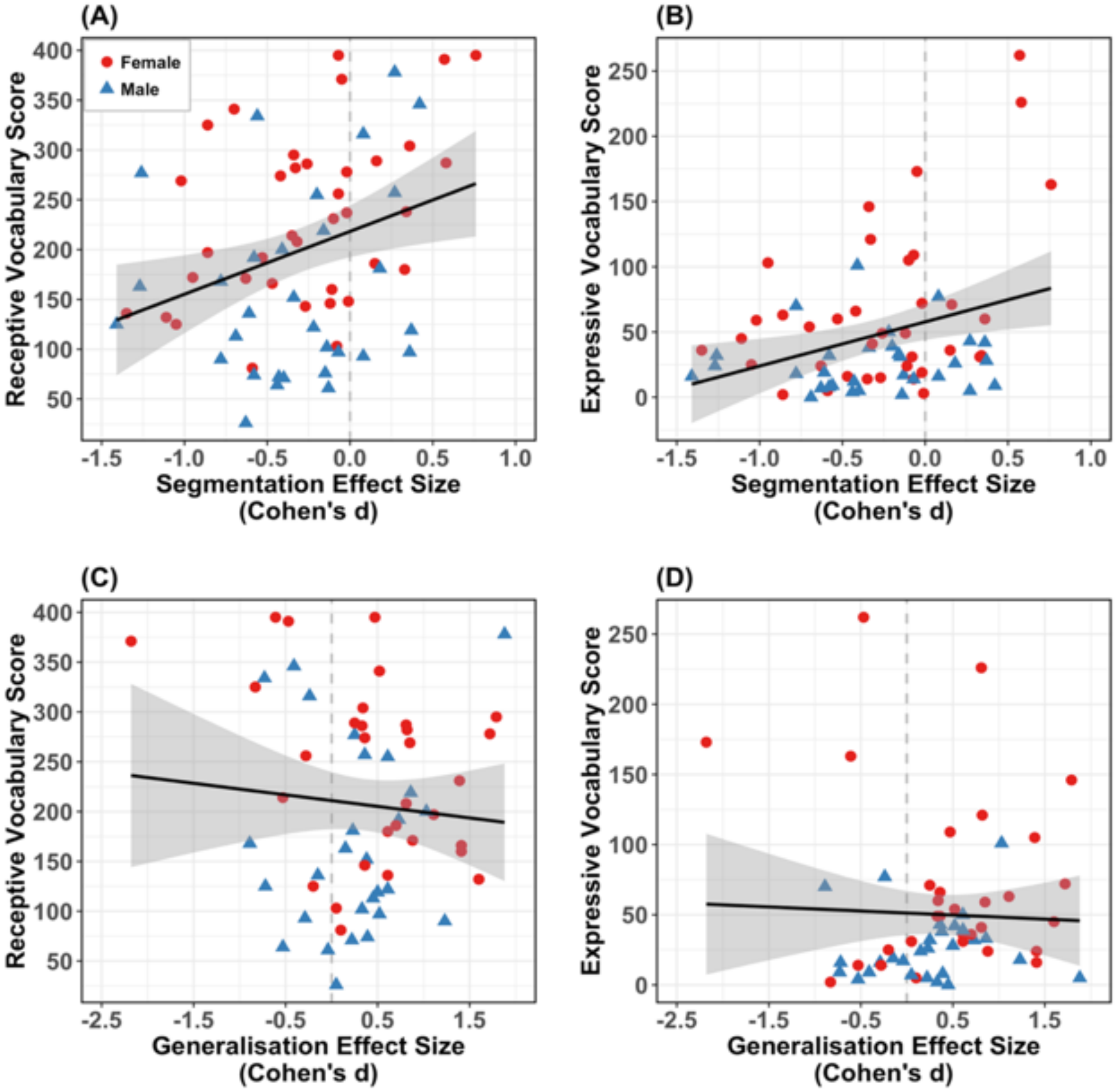

Figure 5. Scatterplots to show the relationship between concurrent vocabulary scores and performance on the segmentation (panel A: receptive; panel B: expressive) and generalisation trials (panel C: receptive; panel D: expressive).

3.2.4.2 Relationship with later language development: Segmentation only

To assess the relationship between segmentation performance and vocabulary acquisition, growth curve analyses (GCA; Mirman, 2014) were performed using *lme4* 1.1.21 (Bates, Mächler, Bolker, & Walker, 2015) in *R* 3.5.2 (R Core Team, 2018). Separate models were fitted to the receptive and expressive vocabulary scores, which were derived from the Lincoln CDI scores taken at 19, 21, 24, 25, 27, and 30 months. As in the previous analyses, the Cohen's d segmentation score was entered as a fixed predictor. To identify the appropriate polynomial order for the age parameter, two separate models were fitted to the data, then compared. The first included age as a centred first-order linear variable, along with the fixed effect of segmentation score. The second entered age as a second-order orthogonal polynomial, in addition to the linear term included in the first model. Both models were fitted with random intercepts for subject, but with no random slopes to maximise comparability.

Model comparison (log-likelihood) indicated a significant difference in model fit ($\chi^2(2)$ = 38.46, p < .001), with the model containing a second order (quadratic) age parameter (AIC = 3601; BIC = 3631) more likely than the alternative with a first-order linear term for age (AIC = 3636; BIC = 3658). Thus, our GCA model contained an orthogonal quadratic age parameter crossed with the fixed effect of segmentation score. The model was fitted with the maximal random effects structure supported by the data (Barr, Scheepers, Levy & Tily, 2013), which included the random intercept of subject, without random slopes. Confirmatory tests were performed using log likelihood-ratios via sequential model decomposition (Bates et al., 2015) with bootstrapped simulations (*R* = 10000) to obtain 95% CIs and p-values for model estimates (Luke, 2017). The marginal and conditional pseudo-R^2 are also reported for the growth curve model, which represent the proportion of the variance explained by fixed effects alone and the full model, respectively (e.g., Nakagawa, Johnson, & Schielzeth, 2017).

For receptive vocabulary, the GCA demonstrated a significant linear increase in scores across development (β = 224.07 [198.34, 249.77], *SE* = 13.12, χ^2 = 127.76, p < .001), but also a quadratic shift in this slope over time (β = -40.34 [-53.37, -27.21], *SE* = 6.67, χ^2 = 54.31, p < .001, see Figure 6). While segmentation ability did not have a significant main effect on the intercept (β = 59.47 [9.75, 109.29], *SE* = 25.39, χ^2 = 2.13, p = .155), it did interact with the linear term of age (β = -47.33 [-90.74, -3.72], *SE* = 22.2, χ^2 = 4.55, p = .039), suggesting that the predictive effect of statistical segmentation ability on vocabulary decreased over development. The fixed effects accounted for 46.85% of the variance in the data, increasing to 93.36% with the inclusion of the random effects (R_m^2 = 0.47; R_c^2 = 0.93).

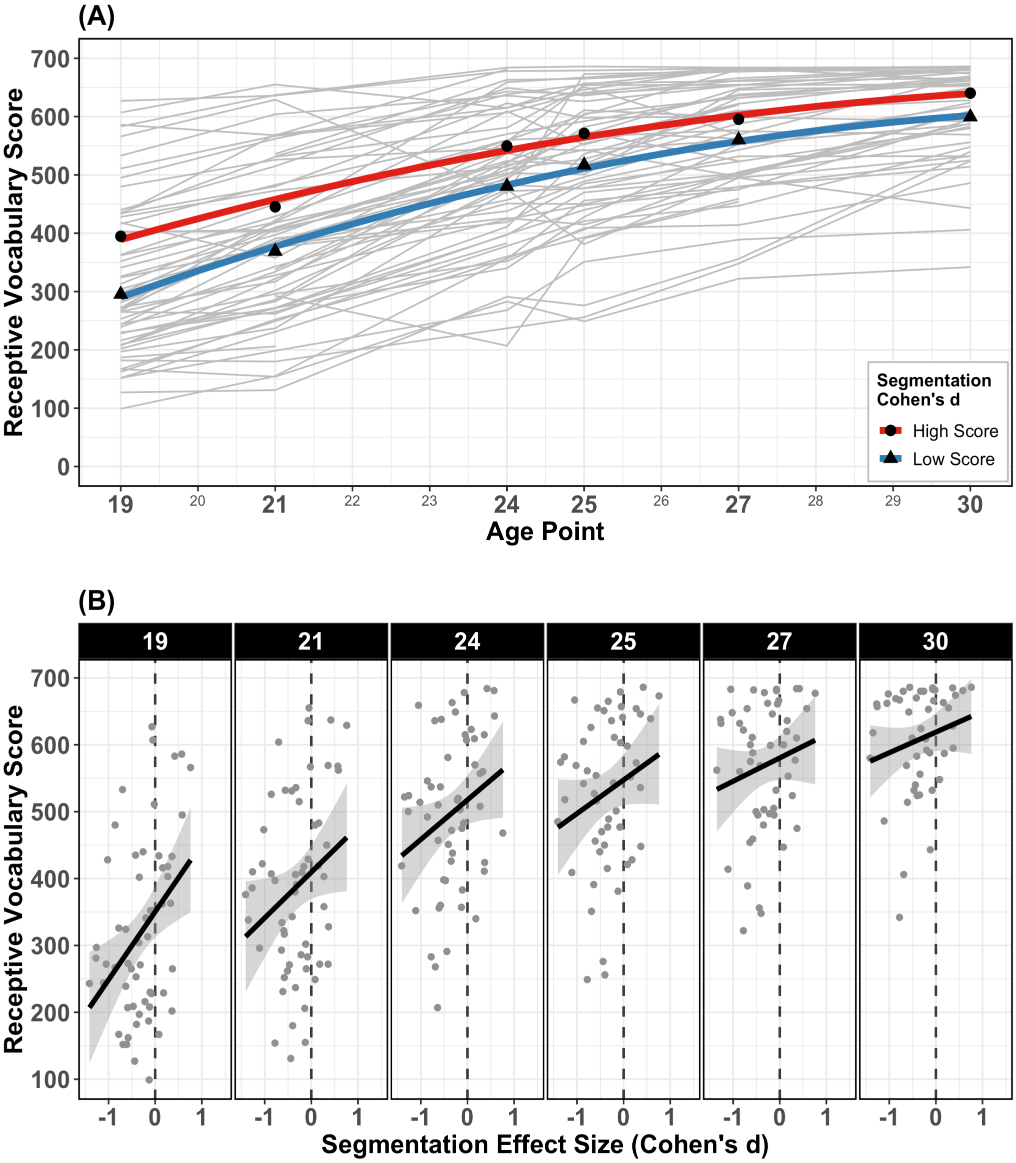

Figure 6. The relationship between segmentation (Cohen's d) at 17 months and receptive vocabulary scores over time (19-30 months). Panel A maps the trajectory of vocabulary development for individual participants (given in grey) and for participants providing high (red; > 0; novelty preference) versus low (blue; < 0; familiarity preference) segmentation scores. Panel B depicts the relationship between segmentation and receptive vocabulary scores at each individual time point.

Similarly, for expressive vocabulary, GCA model fit was significantly improved with the addition of a second-order quadratic age term ($\chi^2(8) = 22.86$, $p < .001$, AIC = 3821, BIC = 3851), compared to a model with only a first-order linear term (AIC = 3840; BIC = 3863).

There was a significant linear increase in scores across development (β = 384.07 [346.15, 422.04], *SE* = 19.36, χ^2 = 137.55, *p* < .001) and a quadratic change in this slope over time (β = -46.86 [-66.31, -27.26], *SE* = 9.96, χ^2 = 34.03, *p* < .001, see Figure 7). Unlike for receptive vocabulary, segmentation ability had a significant positive effect on the intercept (β = 70.56 [6.44, 134.85], *SE* = 32.76, χ^2 = 4.55, *p* = .039), with children who demonstrate larger segmentation effects at 17 months having larger expressive vocabularies. Segmentation scores showed no interaction with the linear (β = -4.8 [-67.91, 58.75], *SE* = 32.31, χ^2 = 0.03, *p* = .856) or quadratic terms of age (β = 12.14 [-19.90, 44.76], *SE* = 16.5, χ^2 = 0.55, *p* = .464). The model explained 56.04% of the variance in the data without the random effects, and 92.81% when they were included (R_m^2 = 0.56; R_c^2 = 0.93).

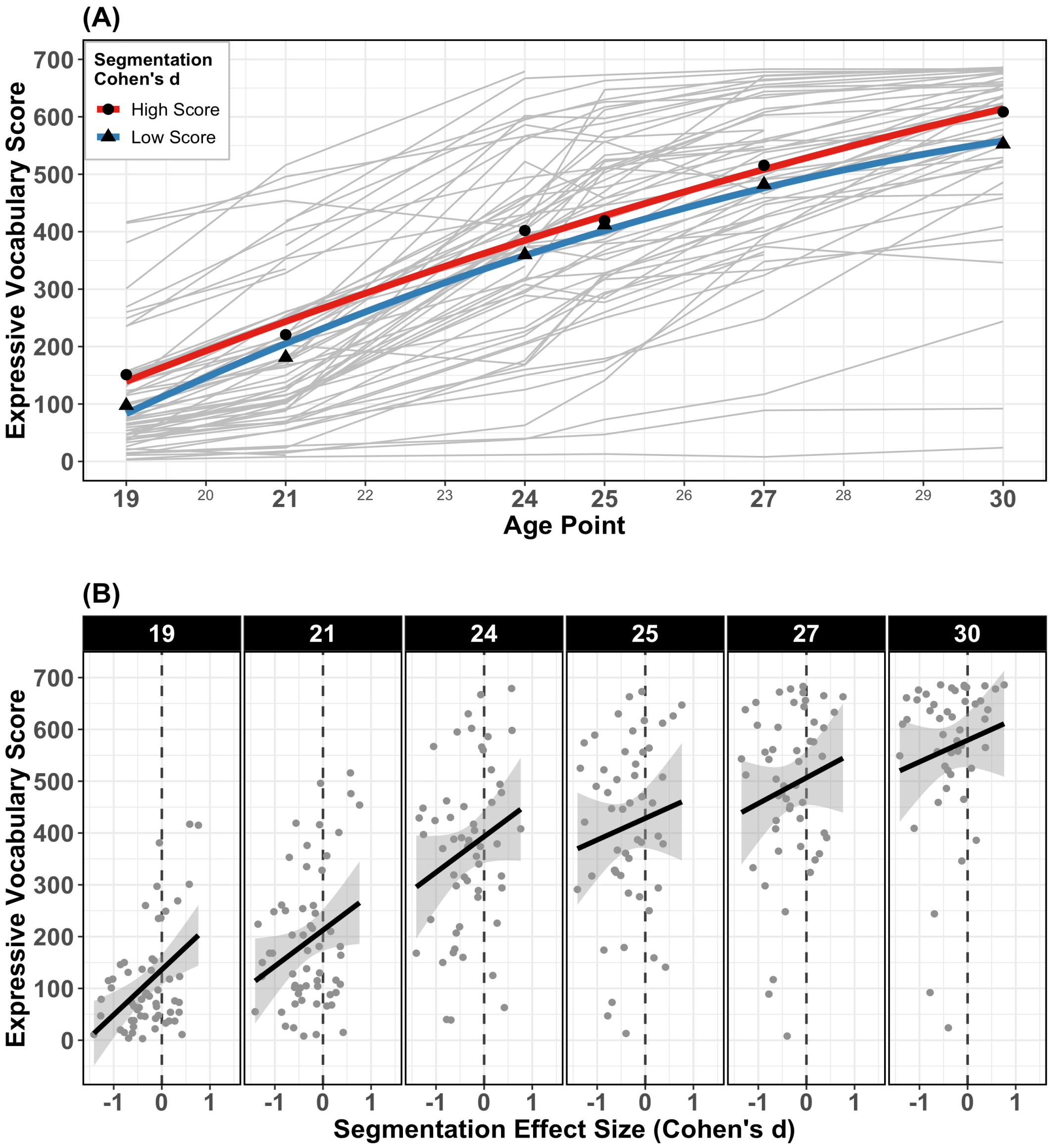

Figure 7. The relationship between segmentation scores (Cohen's d) at 17 months and expressive vocabulary scores over time. Panel A maps the trajectory of vocabulary development for individual participants (given in grey) and for participants providing high (red; > 0; novelty preference) versus low (blue; < 0; familiarity preference) segmentation scores. Panel B depicts the relationship between segmentation and expressive vocabulary scores at each time point.

Although segmentation abilities were shown to predict both expressive and receptive vocabulary, the GCAs suggest that the nature of this relationship may differ across these two measures over time. However, this difference could be due to limitations in the CDI scales:

The relationship between segmentation and receptive vocabulary is seen to plateau around ceiling from 25 months onward, but for expressive vocabulary this is not the case. Thus, it is possible that the receptive measure was unable to capture variance in vocabulary at these later time-points. This is illustrated in Figure A1 (see appendices); receptive vocabulary scores appear to be normally distributed up to ~25 months, but are negatively skewed thereafter.

Additional exploratory analyses were conducted to establish the specific age points at which segmentation ability statistically predicted receptive vocabulary size. Separate bootstrapped multiple regression models were fitted to the data at 24, 25, 27, and 30 months. These models contained the fixed effect of segmentation ability (there was no within-subject random variance to control for). The results (given in tables A1a and A1b, see appendices) suggest that segmentation significantly predicted receptive vocabulary at 24 months (*p* = .031) and marginally at 25 months (*p* = .053), but not at 27 and 30 months.

The prior GCA model was thus refitted using only the receptive vocabulary scores from 19 to 25 months. Adding a second-order polynomial term for age did not improve model fit (χ^2(8) = .87, p = .649). As in the previous models, vocabulary linearly increased with age (β = 31.79, [27.66, 35.95], *SE* = 2.11, χ^2 = 113.05, p < .001). Importantly, segmentation ability positively predicted vocabulary size (β = 55.91, [1.40, 110.57], *SE* = 27.85, χ^2 = 4.05, p = .051); children with larger segmentation scores at 17 months had superior receptive vocabularies across these time points. There was no significant interaction between age and segmentation score (β = -6.87 [-13.90, 0.2], *SE* = 3.6, χ^2 = 3.57, p = .064), suggesting the predictive effect of segmentation on receptive vocabulary size was consistent between 19 and 25 months. Fixed effects accounted for 34.81% of the variance, whereas the entire model (with maximal random effects) explained 93.9% of the variance (R^2_m = 0.35; R^2_c = 0.94).

4. Discussion

We examined 17-month-old infants' ability to learn statistically-defined non-adjacent dependencies from continuous speech, to shed light on whether word segmentation and structure learning may proceed together during language acquisition, from distributional statistics alone (i.e., in the absence of additional cues) - as has recently been demonstrated for adults (Frost & Monaghan, 2016). Demonstrating that infants share this same capacity for statistical learning of words and structure would provide critical insight into the nature of the processes that may underlie these tasks in natural language learning, and the time-course in which they may operate. Crucially, we also investigated the way that infants' statistical learning abilities related to their concurrent natural language skills, and subsequent language development, to help shape our understanding of the way in which statistical learning skills may serve (or be served by) language learning more broadly.

We expected to show that infants could compute over the statistical properties of the speech in order to segment it into individual items (Marchetto & Bonatti, 2013; 2015). Analysis of the segmentation data revealed a significant effect of word type on infants' looking times, indicating that infants attended differently to words and part-words at test. This suggests that infants could indeed compute over the statistical properties of the speech to segment it into word candidates, which they could distinguish from competitor items. These data therefore replicated the finding that infants can segment a continuous stream of artificial speech on the basis of the statistical information contained within the input (e.g., Saffran et al., 1996, Aslin et al., 1998). Further, these data provide critical support for prior demonstrations of infants' ability to do so by computing over non-adjacent, as well as adjacent, statistics (Marchetto & Bonatti 2013; 2015; see also e.g., Frost & Monaghan, 2016; Peña et al., 2002; Perruchet et al., 2004 for demonstrations of this in adults).

In previous studies of morphosyntactic (within-word) non-adjacent dependency learning, segmentation was typically assessed with word and non-word comparisons (where non-words were sequences that had not occurred during habituation), or with comparisons that tested preferences for words and rule-words together, investigating word- and structure-learning simultaneously (Marchetto & Bonatti, 2013; 2015). Here, we tested each task in isolation, and increased the difficulty of the segmentation task by using words and *part-words* - statistical competitors comprising the end of one word and the start of another (rather than random combinations of syllables; see e.g., Saffran et al., 1996). Using this more difficult and more robust assessment, we confirmed that infants could segment speech by computing over non-adjacent statistical regularities.

We note that the group-level looking preference observed for the segmentation task is different to that observed by Marchetto and Bonatti (2015), with higher mean looking times for *word* than *part-word* trials. As this study used a fixed trial order (see section 2.4), we do not make any inferences about this preference – and instead draw upon our LMER analyses that take trial order into account. Nevertheless, we note that the directional difference observed between Marchetto and Bonatti's (2015) work and our conceptual replication could be due to a number of possibilities, including trial order, exposure duration (Endress & Bonatti, 2007, found increasing habituation to words over part-words and generalised words with longer exposure), the different types of test-pair comparisons used, and potential overall differences related to participants' linguistic maturity at the group level (see Hunter & Ames, 1988) – perhaps due to the linguistic knowledge that infants bring to the task, as shaped by their prior experience with relevant language structure (our infants were acquiring English, which is morphologically poor). Future studies which counterbalance presentation order and examine infants' learning cross-linguistically will be key to disentangling these possibilities.

Critically, infants' segmentation performance (indexed by Cohen's d) was found to correlate significantly with their concurrent vocabulary size, both for receptive and expressive vocabulary, providing further evidence that infants' capacity for speech segmentation in the laboratory relates meaningfully to their real-word language skills (Junge et al., 2012; Kidd et al., 2018; Newman et al., 2006; Newman et al., 2016; Singh et al., 2012), and extending this finding to statistical segmentation of non-adjacent dependencies. These data indicate that infant's statistical language learning abilities may shape, or be shaped by, infant's language proficiency (see also Lany, 2014, and Lany and Shoaib, 2019). This demonstration that statistical learning of non-adjacencies supports segmentation of an artificial language stream serves as striking evidence that such artificial language learning studies are probing key mechanisms in natural language development.

Of particular note is the direction of the learning effects on the segmentation task, and the way that the polarity of Cohen's d scores related to infants' CDI scores; infants with lower CDI scores demonstrated a familiarity effect (preferring words), whereas infants with higher CDI scores demonstrated a novelty effect (preferring part-words). This difference is suggestive of a maturational preference-switch (from familiarity to novelty), similar to that demonstrated in the ERP segmentation literature (Kidd et al., 2018). This result is in line with the prior suggestion that infants' looking preferences are dynamic, with directional switches resulting from differences in levels of stimulus encoding (see e.g., Houston-Price & Nakai, (2004); Hunter & Ames (1988), and see e.g., Jusczyk & Aslin (1995) and Saffran et al. (1996) for a demonstration of preferential differences on speech segmentation tasks that could (at least in part) be due to differences in exposure and stimulus encoding).

The results from the growth curve analyses indicate that the relationship observed between statistical segmentation ability and vocabulary size at the time of testing persists over development, with segmentation performance significantly statistically predicting receptive vocabulary up to 24 months, and expressive vocabulary up to 30 months. Thus, we propose that infants' statistical learning ability may be an informative predictor of their vocabulary size at a later point in development - possibly even over a year later. However, in the study at hand, segmentation performance is not seen to positively influence the *rate* of vocabulary acquisition - with no apparent relationship between segmentation scores and growth for expressive vocabulary (i.e., a stronger preference on the segmentation task did not predict *faster* learning). For receptive vocabulary, there was a negative relationship between segmentation scores and growth, which is likely due to receptive scores reaching ceiling at the later time points.

The lack of relationship between segmentation ability and vocabulary growth could be interpreted in a number of ways. One possibility is that individual differences in statistical learning are unrelated to individual differences in vocabulary acquisition. However, this is unlikely; the data reveal a strong relationship between statistical learning and concurrent vocabulary, and the GCA show a significant effect on the intercept over development - indicating that these abilities are indeed related. A second possibility is that both differences in statistical learning and differences in vocabulary acquisition are due to another underlying factor not assessed here, for instance individual differences in neural maturation (general cognitive ability), speed of processing, or possible differences arising from variation in the socioeconomic background of infants (see e.g., Schwab & Lew-Williams, 2018). Testing the possible influence of additional variables would require assessing a broader array of cognitive skills, and factoring infants' performance on these additional tasks into the analyses.

A third alternative is that differences in statistical learning contribute to differences in the rate of vocabulary acquisition at the earliest stages of language development, so are not captured here. It may be the case that infants' strategies for segmentation change over development, with infants relying more on statistical learning to develop their early lexicon, then incorporating other strategies when they become available, or when infants reach a certain level of proficiency (see e.g., Conway et al. (2010), and Frost, Monaghan, & Christiansen (2019) for evidence that learners may make use of both bottom-up and top-down strategies for speech segmentation). The notion of a developmental shift in infants' speech segmentation strategy is not new; there is much research to suggest that early segmentation is stress-based, before infants turn to the statistical properties of the input (e.g., Johnson & Juscyzk, 2001;

Johnson & Seidl, 2009; Thiessen & Saffran, 2003). It is conceivable that infants then adapt their segmentation strategy further upon reaching a certain level of maturity; for instance, by drawing on representations for highly familiar items to help identify word boundaries for neighbouring items (e.g., Bortfeld, Morgan, Golinkoff, & Rathbun 2005). An early advantage of statistical learning for vocabulary acquisition could explain why children with good statistical learning abilities have bigger vocabularies in the study at hand, though further research examining the relationship between statistical learning and vocabulary growth earlier in development is required to test these claims.

Contrary to prior suggestions (Marchetto & Bonatti, 2013; 2015), there was some evidence that infants could generalise the non-adjacent dependencies contained within continuous speech to grammatically consistent but previously unseen sequences (trained dependencies with a novel intervening syllable). Previous studies have demonstrated generalisation only when pauses separated sequences containing the dependencies (e.g., Marchetto & Bonatti, 2013; 2015). The apparent necessity of this pause has been interpreted as requiring speech segmentation to be resolved before generalisation over the grammatical structure can occur. However, in the current study we showed that this additional pause cue was not necessary, and that generalisation could occur in the same brief learning period as segmentation, from the same input.

There are two key possible explanations for the generalisation effects seen here. The first is that 17-month-old infants are able to generalise non-adjacent dependencies under the right testing circumstances, with conflicting use of syllables across test items preventing participants from distinguishing between rule-words and part-words in prior studies (e.g., Marchetto & Bonatti, 2015, see Frost & Monaghan, 2016). Thus, the data could indicate that without this conflicting information, generalisation of the non-adjacencies can be observed in infancy in the absence of additional cues to the language structure (e.g., Mueller et al., 2008, 2010) – perhaps proceeding together with speech segmentation.

However, an alternative possibility is that generalisation performance could be a product of test order, with infants acquiring within-word structure over the course of the tasks rather than during familiarisation; infants heard segmentation trials first, and so received additional exposure to the words from the speech stream ahead of the generalisation task. Since words were presented in isolation on the segmentation trials, infants were not necessarily required to learn about the structure of those words *while* segmenting them from speech in order to succeed on this task. This is unlikely to explain infants' generalisation performance entirely, though, as infants also received equivalent exposure to part-words during the segmentation testing - meaning the segmentation task may have strengthened the child's representations of both words and part-words (which formed non-words) to a similar degree. Nevertheless, it remains a possibility that completing this task first could have influenced subsequent generalisation performance. Future studies addressing generalisation immediately after training (i.e., without the segmentation task) will enable us to firmly disentangle infants' capacity for generalisation from possible effects of task order. Follow-up studies with and without inter-item pauses in the speech stream will also permit a more thorough investigation of infant's capacity to learn structure from speech.

Finding evidence for both segmentation and generalisation could suggest that both tasks may be supported similarly by the same statistical properties of the input. However, the relationship between these tasks, and their differing links to vocabulary development, does not permit us to confirm that the same statistical operations are applying to both tasks (Frost & Monaghan, 2017). We note, though, that the relative draw toward familiar versus novel items may have been somewhat different across these tasks, which could restrict the degree to which these scores could be compared directly; whereas the segmentation task involves a straightforward familiarity (words) versus novelty (part-words) comparison, the generalisation task is perhaps more complex, and could be interpreted as pitting less novelty (rule-words) against more novelty (non-words).

Nevertheless, the results indicate some distinctions between these tasks. First, whereas both tasks individually demonstrate learning, the correlation between segmentation and generalisation performance was not significant, meaning there was no observable relationship between performance on these tasks. Second, while the relationship between vocabulary size and segmentation performance was significant, for generalisation this was not the case; CDI scores did not correlate with performance. This is not the same as a dissociation, and the lack of correlation could have been due to the lower sensitivity of the generalisation task compared to the segmentation task (cf. mean and SD of estimates in Tables 1 and 2), rather than an absence of a relationship altogether. The alternative - that segmentation and generalisation are radically different types of tasks (Marchetto & Bonatti, 2015; Peña et al., 2002) - would predict that only vocabulary scores should relate to segmentation performance here, whereas generalisation performance ought only to relate to tasks associated with distinct grammatical processing. The present results may be seen to better align with the latter, however future tests of these children's grammatical processing abilities would be required to address this directly.

In sum, this study provides further evidence that infants can segment speech by computing over the statistical properties of the input (in this case, non-adjacent dependencies). We find evidence to suggest that children can also detect non-adjacent dependency structure in continuous speech, and generalise this to novel consistent items, though further research is required to establish this conclusively. Crucially, we have shown that laboratory-based studies of children's language learning, in terms of abstract word segmentation from non-adjacent structures in continuous artificial speech, have real-world counterparts in children's language development. Furthermore, we have shown that an individual differences approach to interpreting the effect size on these artificial language learning tasks differentiates children who present with familiarity and novelty preferences, with the direction of the effect corresponding with children's vocabulary size (but not with their vocabulary *growth*). This insight into individual differences in performance shows that such variation is meaningful, rather than noise, and contributes to further interpretation of novelty and familiarity preferences with respect to language maturation.

References

Alcock, K. J., Meints, K., & Rowland, C. F., (2020). *The UK Communicative Development Inventory: Words and Gestures*. Guilford, UK: J&R Press Ltd.

Aslin, R. N., Saffran, J., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9,* 321–324. doi: 10.1111/1467-9280.00063

Aslin, R. N., Woodward, J., LaMendola, N., & Bever, T. (1996). Models of word segmentation in fluent maternal speech to infants. In J. Morgan and K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.* Mahwah, NJ: Lawrence Erlbaum.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. doi:10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1–48. doi:10/gcrnkw

Bion, R. A. H., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word-object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition, 126*, 39-53. doi:10.1016/j.cognition.2012.08.008

Black, A. W., Taylor, P., & Caley, R. (1990). The festival speech synthesis system. United Kingdom: Centre for Speech Technology Research (CSTR), University of Edinburgh. Retrieved from http://www.cstr.ed.ac.uk/projects/festival.html.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science, 16,* 298–304. doi:10.1111/j.0956-7976.2005.01531.x

Christiansen, M. H, Conway C. M., & Onnis L. (2012). Similar neural correlates for language and sequential learning: Evidence from event-related brain potentials. *Language and Cognitive Processes, 27,* 231–256. doi:10.1080/01690965.2011.606666

Cohen, J. (1992). A power primer. *Psychological bulletin, 112*(1), 155. doi:10/as7

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition, 114*, 356-371. doi:10.1016/j.cognition.2009.10.009

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. doi:10.3389/fpsyg.2017.01482

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures*.* *Child Development, 85*(4), 1130-1145. doi:10.1111/cdev.12193

Endress, A. D. & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition, 105*, 247-299.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., Bates, E. (2007). MacArthur-Bates Communicative Development Inventories: User's guide and technical manual (2nd ed.). Baltimore, MD: Brookes.

Frost, R. L. A., Isbilen, E. S., Christiansen, M. H. & Monaghan, P. (2019). Testing the limits of non-adjacent dependency learning: Statistical segmentation and generalization across domains. In A.K. Goel, C.M. Seifert, & C. Freksa (Eds.) *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. Montreal, QB: Cognitive Science Society.

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.

Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2019). Mark my words: high frequency marker words impact early stages of language learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*. doi:10.1037/xlm0000683

Gerken, L. A. (2006). Decisions, decisions: infant language learning when multiple generalisations are possible. *Cognition, 98*, B67–B74. doi:10.1016/j.cognition.2005.03.003

Gerken, L. A. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition, 115*, 362-366. doi:10.1016/j.cognition.2010.01.006

Gómez, R. L. (2002). Variability and detection of invariant structure, *Psychological Science, 13*(5) 431-436. doi:10.1111/1467-9280.00476

Gómez, R. L., Bootzin, R. R, & Nadel, L. (2006). Naps Promote Abstraction in Language-Learning Infants. *Psychological Science, 17*(8), 670-674.

Gómez, R., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109–135. doi:10.1016/S0010-0277(99)00003-7

Gómez, R. L. & Maye, J. (2005). The developmental trajectory of non-adjacent dependency learning. *Infancy, 7(2)*, 183-206. doi:10.1207/s15327078in0702_4

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development, 13*, 341-348. doi:10.1002/icd.364

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. Advances in Infancy Research, 5, 69-95.

Isbilen, E S., Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2018), Bridging artificial and natural language learning: Chunk-based computations in learning and generalizing statistical structure. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds*.), Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1856-1861). Austin, TX.

Johnson, E. K. & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: when speech cues count more than statistics. Journal of Memory and Language, 44(4), 548-567. doi:10.1006/jmla.2000.2755

Johnson, E. K., & Seidl, A. H. (2009). At 11 months, prosody still outranks statistics. Developmental Science, 12(1), 131-141. doi:10.1111/j.1467-7687.2008.00740.x

Junge, C., Kooijman, V., Hagoort, P., & Cutler, A. (2012). Rapid recognition at 10 months as a predictor of language development, *Developmental Science, 15(4),* 463-473. doi:10.1111/j.1467-7687.2012.1144.x

Jusczyk, P. & Aslin, R. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29,* 1-23. doi:10.1006/cogp.1995.1010

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. doi:10.1016/0163-6383(95)90012-8

Kidd, E. (2012). Implicit statistical learning is directly associated with the acquisition of syntax. *Developmental Psychology, 48*(1), 171-184. doi:10.1037/a0025405

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development, 87,* 184–193. doi:10.1111/cdev.12461

Kidd, E., Junge, C., Spokes, T., Morrison, L., & Cutler, A. (2018). Individual differences in infant speech segmentation: Achieving the lexical shift. *Infancy, 23*(6), 770-794. doi:10.1111/infa.12256

Lany, J. (2014). Judging words by their covers and the company they keep: Probablistic cues support word learning. *Child Development, 85*(4), 1727-1739. doi:10.1111/cdev.12199

Lany, J. & Gómez, R. L. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science, 19(12),* 1247 – 1252. doi:10.1111/j.1467-9280.2008.02233.x

Lany, J. & Gómez, R. L., & Gerken, L. (2007). The role of prior experience in language acquisition, *Cognitive Science, 31,* 481-507. doi:10.1080/15326900701326584

Lany, J. & Shoaib, A. (2019). Individual differences in non-adjacent statistical dependency learning in infants. *Journal of Child Language*, *13,* 1-25.

Lashley, K. S. (1951). The problem of serial order in behavior. In L.A. Jeffress (Ed.), *Cerebral Mechanisms in Behavior* (pp. 112- 146). New York: Wiley.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*(4), 1494–1502. doi:10/gbsd4m

Marchetto, E. & Bonatti, L. L. (2013). Words and possible words in early language acquisition. *Cognitive Psychology, 67*(3), 130–50. 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

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven month-old infants. *Science, 283,* 77-80. doi:10.1126/science.283.5398.77

Meints, K., & Fletcher, K. L. (2001). Toddler communicative development inventory, a UK adaptation of the MacArthur-Bates communicative development inventories. University of Lincoln Babylab.

Mirman, D. (2014). Growth curve analysis and visualization using R. USA: CRC Press.

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science, 2*, 138–153. doi:10.1111/j.1756-8765.2009.01072.x

Molfese, D. L. (2000). Predicting dyslexia at 8 years of age using neonatal brain responses. *Brain and Language, 72* (3), 238-245. doi:10.1006/brln.2000.2287

Molfese, D. L., & Molfese, V. J. (1985). Electrophysiological indices of auditory discrimination in newborn infants. *Infant Behavior and Development, 8,* 197–211. doi:10.1016/S0163-6383(85)80006-0

Molfese, D. L. & Molfese, V. J. (1997). Discrimination of language skills at five years of age using event-related potentials recorded at birth. *Developmental Neuropsychology, 13*(2), 135-156. doi:10.1080/87565649709540674

Monaghan, P. (2017). Canalization of language structure from environmental constraints: A computational model of word learning from multiple cues. *Topics in Cognitive Science, 9,* 21-34. doi:10.1111/tops.12239

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The Phonological Distributional coherence Hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology, 55,* 259-305. doi:10.1016/j.cogpsych.2006.12.001

Mueller, J. L., Bahlmann, J., & Friederici, A. D. (2008). The role of pause cues in language learning: The emergence of event-related potentials related to sequence processing. *Journal of Cognitive Neuroscience, 20,* 892–905. doi:10.1162/jocn.2008.20511

Mueller, J. L., Bahlmann, J., & Friederici, A. D. (2010). Learnability of embedded syntactic structures depends on prosodic cues. *Cognitive Science, 34*, 338–349. doi:10.1111/j.1551-6709.2009.01093.x

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*(134), 20170213. doi:10/gddpnq

Newman, R. S., Bernstein Ratner, N., Jusczyk, A. M., Jusczyk, P. W., & Dow, K. A. (2006). Infants' early ability to segment the conversational speech signal predicts later language development. *Developmental Psychology, 42,* 337–367. doi:10.1037/0012-1649.42.4.643

Newman, R. S., Rowe, M. & Bernstein Ratner, N. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language, 43*(5), 1158-1173. doi:10.1017/S0305000915000446

Newport, E. L. & Aslin, R. (2004). Learning at a distance: Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48(*2), 127-162. doi:10.1016/S0010-0285(03)00128-2

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in speech processing. *Journal of Memory and Language, 53,* 225–237. doi:10.1016/j.jml.2005.02.011

Panter A. T., Tanaka J. S., & Wellens T. R. (1992). The Psychometrics of Order Effects. In: Schwarz N., Sudman S. (eds) *Context Effects in Social and Psychological Research*. Springer, New York, NY.

Peña, M., Bonatti, L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298*, 604-607. doi:10.1126/science.1072901

Pelucchi, B., Hay, J. F., Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development, 80*(3), 674-685. doi:10.1111/j.1467-8624.2009.01290.x

Perruchet, P., Tyler, M. D., Galland, N., & Peereman, R. (2004). Learning non- adjacent dependencies: No need for algebraic-like computations. *Journal of Experimental Psychology, 133*(4), 573-583). doi:10.1037/0096-3445.133.4.573

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 Psycholoy, 115, 101238.

R Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Redington, M. & Chater, N. (1997). Probabilistic and distributional approaches to language acquisition. *Trends in Cognitive Science, 1(7),* 273-281. doi:10.1016/S1364-6613(97)01081-4

Rivera-Gaxiola, M., Klarman, L., Garcia-Sierra, A., & Kuhl, P. K. (2005). Neural patterns to speech and vocabulary growth in American infants. *NeuroReport (Developmental Neuroscience), 16,* 495–498.

Rowland, C. F., Bidgood, A., Durrant, S., Peter, M., & Pine, J. M. (unpub.). The Language 0-5 Project. Unpublished manuscript, University of Liverpool. doi:10.17605/OSF.IO/KAU5F.

Rubenstein, H. (1973). Language and probability. In G. A. Miller (Ed.), Communication, language, and meaning: Psychological perspectives (pp. 185-195). New York: Basic Books, Inc.

Saffran, J. R, Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928. doi:10.1126/science.274.5294.1926

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*(2), 101-105. doi:10.1111/j.1467-9280.1997.tb00690.x

Schwab, J. F., & Lew-Williams, C. (2016). Language learning, socioeconomic status, and child-directed speech. Wiley interdisciplinary reviews. Cognitive science, 7(4), 264–275.

Singh, L., Reznick, J. R., & Xuehua, L. (2012). Infant word segmentation and childhood vocabulary development. *Developmental Science, 15*, 482–495. doi:10.1111/j.1467-7687.2012.01141.x

Thiessen, E.D., & Saffran, J.R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39,* 706–716. doi:10.1037/0012-1649.39.4.706

Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2004). Speech perception in infancy predicts language development in the second year of life. *Child Development, 75*, 1067–1084. doi:10.1111/j.1467-8624.2004.00726.

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/L008955/1].

Appendices

Table A1a. Summary of the bootstrapped regression models conducted for segmentation scores and vocabulary at 24, 25, 27, and 30 months

Age	Term	*β*	*SE*	*t*	*p*
24 Months	*Intercept*	517.61 [484.39, 550.93]	16.98	30.49	< .001
	Segmentation	59.24 [4.38, 111.69]	27.37	2.16	.031
25 Months	*Intercept*	547.45 [514.79, 581.23]	16.95	32.30	< .001
	Segmentation	50.3 [-0.4, 101.57]	26.01	1.93	.053
27 Months	*Intercept*	580.33 [554.15, 607.93]	13.72	42.30	< .001
	Segmentation	34.88 [-16.3, 85.14]	25.88	1.35	.178
30 Months	*Intercept*	618.94 [598.3, 640.4]	10.74	57.63	< .001
	Segmentation	30.87 [-10.07, 70.8]	20.63	1.50	.135

Table A1b. Summary of fit for the bootstrapped regression models conducted for segmentation scores and vocabulary at 24, 25, 27, and 30 months

Age	Model Fit
24 Months	F(1, 52) = 3.97 [-5.06, 10.93], p = .052, R^2 = 0.071
25 Months	F(1, 49) = 2.73 [-4.11, 7.97], p = .105, R^2 = 0.053
27 Months	F(1, 49) = 1.6 [-4.6, 6.13], p = .212, R^2 = 0.032
30 Months	F(1, 45) = 1.85[--3.99, 6.24], p = .181, R^2 = 0.039

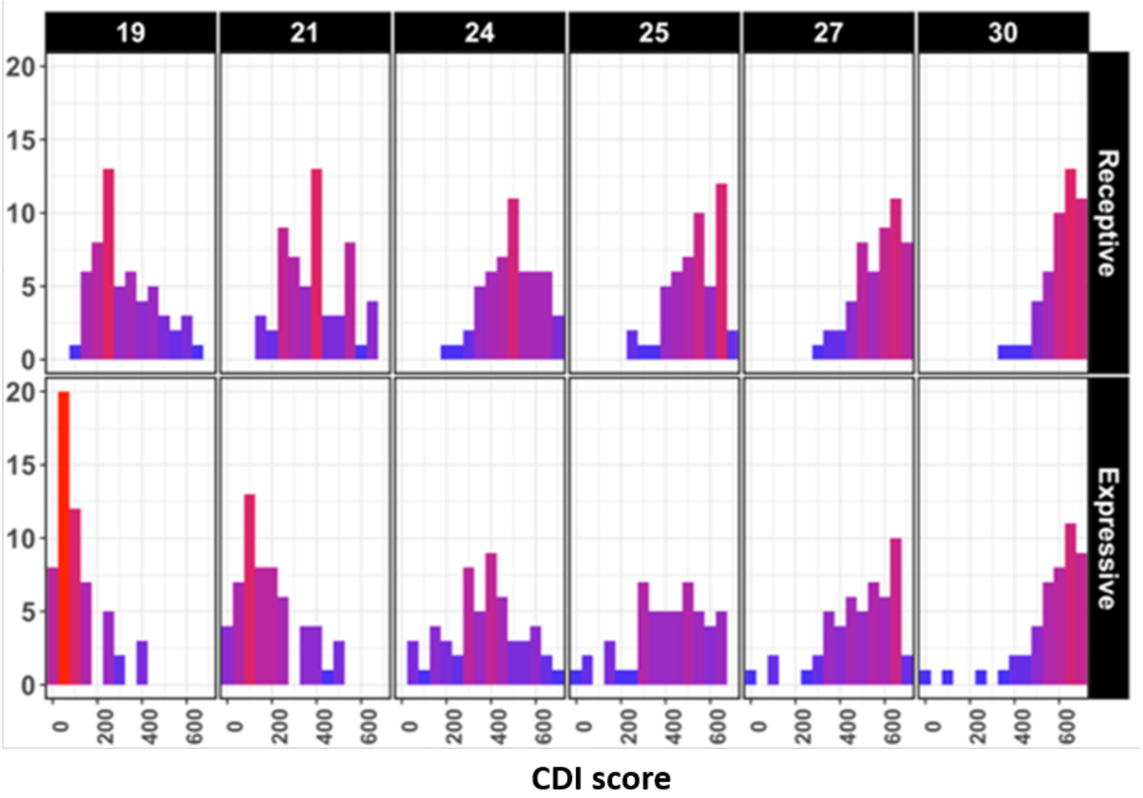

Figure A1. Histograms depicting the distribution captured by the CDI measures for expressive and receptive vocabulary scores across development.