L A N G U A G E
D E V E L O P M E N T
R E S E A R C H

An Open Science Journal

Volume 3 | Issue 1 | December 2023
ISSN 2771-7976

About the journal

Language Development Research: An Open-Science Journal was established in 2020 to meet the field's need for a peer- reviewed journal that is committed to fully open science: LDR charges no fees for readers or authors, and mandates full sharing of materials, data and analysis code. The intended audience is all researchers and professionals with an interest in language development and related fields: first language acquisition; typical and atypical language development; the development of spoken, signed or written languages; second language learning; bi- and multilingualism; artificial language learning; adult psycholinguistics; computational modeling; communication in nonhuman animals etc. The journal is managed by its editorial board and is not owned or published by any public or private company, registered charity or nonprofit organization.

Child Language Data Exchange System

Language Development Research is the official journal of the **TalkBank system**, comprising the CHILDES, PhonBank, HomeBank, FluencyBank, Multilingualism and Clinical banks, the CLAN software (used by hundreds of researchers worldwide to analyze children's spontaneous speech data), and the Info-CHILDES mailing list, the de-facto mailing list for the field of child language development with over 1,600 subscribers.

Diamond Open Access

Language Development Research is published using the Diamond Open Access model (also known as "Platinum" or "Universal" OA). The journal does not charge users for access (e.g., subscription or download fees) or authors for publication (e.g., article processing charges).

Hosting

The **Carnegie Mellon University Library Publishing Service** (LPS) hosts the journal on a Janeway Publishing Platform with its manuscript management system (MMS) used for author submissions.

License

Peer Review and Submissions

All submissions are reviewed by a minimum of two peer reviewers, and one of our Action Editors, all well- established senior researchers, chosen to represent a wide range of theoretical and methodological expertise. Action Editors select peer reviewers based on their expertise and experience in publishing papers in the relevant topic area.

Submissions and Publication Cycle

We invite submissions that meet our criteria for rigour, without regard to the perceived novelty or importance of the findings. We publish general and special-topic articles ("Special Collections") on a rolling basis to ensure rapid, cost-free publication for authors.

Language Development Research is published once a year, in December, with each issue containing the articles produced over the previous 12 months. Individual articles are published online as soon as they are produced. For citation purposes, articles are identified by the year of first publication and digital object identifier (DOI).

Table of Contents

Volume 3, Issue 1, December 2023

Distributional learning of novel visual object categories in children with and without developmental language disorder

Iris Broedelet
Paul Boersma
Judith Rispens
Amsterdam Center for Language and Communication,
University of Amsterdam, the Netherlands

Abstract: It has been proposed that a deficit in statistical learning contributes to problematic language acquisition in children with developmental language disorder (DLD), but at the same time the nature and extent of this relationship is not clear. This paper focuses on the role of statistical learning in lexical-semantic development by investigating visual distributional learning of novel object categories in children with and without DLD and its relation to vocabulary knowledge. Distributional learning is a form of statistical learning and entails the learning of categories based on the frequency distribution of variants in the environment. Fifty children (25 DLD, 25 TD) were tested on a visual distributional learning task. Results indicate that children can learn novel object categories on the basis of distributional information. We did not find evidence for a deficit in visual distributional learning in children with DLD. To investigate whether visual distributional learning ability is related to vocabulary knowledge, the children with DLD were tested on different measures of vocabulary. Phonological processing ability and non-verbal intelligence were taken into account as control variables. Multiple linear regression analyses did not reveal evidence for a relationship between distributional learning and vocabulary in DLD.

Keywords: developmental language disorder; statistical learning; distributional learning; lexical-semantic knowledge.

Corresponding author(s): Iris Broedelet, Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 134, 1012 VB Amsterdam. E-mail: I.R.L.Broedelet@uva.nl.

Citation: Broedelet, I., Boersma, P., & Rispens, J. (2023). Distributional learning of novel visual object categories in children with and without developmental language disorder. *Language Development Research, 3*(1), 1-43, https://doi.org/10.34842/2023.0528.

Introduction

Most children acquire their native language(s) without many major difficulties, but this is different for children with developmental language disorder (henceforth: DLD). These children do not present major neurological deficits, hearing disabilities or low overall intelligence, nor is a lack of language input the underlying problem. DLD occurs in approximately 7% of school-aged children (Bishop, 2006), and the problems often last into adulthood. Social–emotional difficulties occur in this group as well: individuals with DLD have greater risk of depression disorders (Westby, 2019) and even have a lower quality of life compared to typically developing peers (Eadie et al., 2018).

Morphosyntactic impairments are viewed as a hallmark of DLD, while lexical abilities are often seen as a relative strength (e.g. Ullman & Pierpont, 2005). However, there is ample clinical evidence for a disadvantage in lexical skills as well (for reviews: Brackenbury & Pye, 2005; Nation, 2014). Recently, researchers have proposed that an impairment in statistical learning, a learning ability that is important for the discovery of patterns and sequences in sensory input (Siegelman et al., 2017), contributes to the language difficulties in children with DLD (Arciuli & Conway, 2018; Hsu & Bishop, 2010; Saffran, 2018). Experimental results suggest that a deficit in statistical learning (partly) explains lexical deficits (Evans et al., 2009; Mainela-Arnold & Evans, 2014), but the relationship between statistical learning and the development of lexical knowledge, especially lexical-semantic knowledge, requires more investigation.

Distributional learning, which plays a role in the categorization of sensory stimuli such as speech sounds (Maye et al., 2008; Maye et al., 2002) and novel visual objects (Junge et al., 2018) has never been investigated in children with DLD. Categorizing novel visual stimuli might be an important skill that is required when mapping new words to new objects. In our study we aim to investigate if this type of visual distributional learning is affected in children with DLD, and whether this ability relates to different types of lexical(-semantic) knowledge.

Background

Statistical learning deficit hypothesis

Although the main aspect of DLD is problematic language acquisition, children with DLD experience difficulties outside the linguistic domain as well. For example, there is evidence for deficits in motor skills (Sanjeevan & Mainela-Arnold, 2019), working memory (Montgomery et al., 2010), attention (Ebert & Kohnert, 2011) and processing visual information (Collisson et al., 2015). These findings have led to the idea that a deficit in a more general learning mechanism might be at the core of the disorder, as

opposed to an impairment specific to linguistic representations (Arciuli & Conway, 2018; Hsu & Bishop, 2010).

Statistical learning is such a learning mechanism (Siegelman, 2020). Statistical learning underlies the extraction of regularities and patterns from sensory input and has been shown to correlate with or predict language ability in children and adults (Conway et al., 2010; Ellis et al., 2014; Hamrick et al., 2018; Kaufman et al., 2010; Kidd & Arciuli, 2016; Kidd, 2012; Misyak et al., 2010; Newman et al., 2006; Shafto et al., 2012; Spencer et al., 2015).

Results from several studies point towards a disadvantage in different types of statistical learning in individuals with DLD (Evans et al., 2009; Haebig et al., 2017; Hsu & Bishop, 2010; Hsu et al., 2014; Lammertink et al., 2019; Lian, 2017; Lukács & Kemény, 2014; Lum et al., 2014; Mainela-Arnold & Evans, 2014; Obeid et al., 2016; Plante et al., 2002; Tomblin et al., 2007); for a review see Saffran (2018). Please note that null results (Aguilar & Plante, 2014; Noonan, 2018) and even evidence of intact statistical learning in children with DLD (Lammertink et al., 2020) have also been reported. Importantly, several meta-analyses point to a statistical learning deficit in children with DLD (Lammertink et al., 2017; Lum & Conti-Ramsden, 2013; Obeid et al., 2016). Moreover, studies have suggested that statistical learning ability is related to different types of language skills in children with DLD: for example grammatical ability (Hedenius et al., 2011; Misyak et al., 2010; Tomblin et al., 2007) and lexical skills (Evans et al., 2009; Mainela-Arnold & Evans, 2014). Thus, accumulated evidence indicates that children with DLD are compromised in their statistical learning ability, which might (partly) explain their problematic language acquisition.

Lexical difficulties in children with DLD

Children with DLD may have difficulty with several aspects of language acquisition, such as vocabulary, morphology, syntax and phonology, and there is a large amount of heterogeneity within this population (Bishop, 2006; Leonard, 2014). Many studies have focused on morphosyntactic difficulties, for example a child saying she walk instead of she walks. However, these children also show evident difficulties in the development of lexical knowledge (Brackenbury & Pye, 2005; Nation, 2014). Research indicates that lexical difficulties impact social and academic development (Aguilar et al., 2017).

Studies suggest that children with DLD have a smaller vocabulary size and more shallow knowledge of words relative to TD children (McGregor et al., 2013). For example, they make semantic substitutions (confusing towel and blanket) and use more “all-

purpose verbs" like go instead of more specific verbs like run, skip, sail, swim, etc. When naming objects, they are slower and make more phonological and semantic errors (Dockrell et al., 2001; Lahey & Edwards, 1999; Leonard et al., 1983; McGregor et al., 2002; McGregor, 1997). These errors reflect impoverished semantic representations. Dockrell et al. (2003) tested semantic knowledge of children with word-finding difficulties, and found that they provide less accurate definitions of objects and actions: their definitions often contained less information about the semantic category of an object, and more perceptual and redundant information compared to TD children. Moreover, compared to controls, children with DLD provide poor, incomplete definitions of common words (Mainela-Arnold et al., 2010; Marinellie & Johnson, 2002), and provide fewer semantic details in drawings (McGregor & Appel, 2002; McGregor et al., 2002).

On word association tasks, which are viewed as a measure of lexical-semantic organization, children with DLD produce fewer semantically related words than TD peers (Drljan & Vuković, 2019; McGregor et al., 2012; Sandgren et al., 2020; Sheng & McGregor, 2010). A less efficient lexical organization could have a negative effect on subsequent vocabulary development (Beckage et al., 2010). Finally, children with DLD also show difficulties on word learning tasks, both with learning phonological and semantic properties of words (Alt & Plante, 2006; Kan & Windsor, 2010; Nash & Donaldson, 2005) and fast mapping (Haebig et al., 2017; Kapa & Erikson, 2020).

Thus, children with DLD have lexical difficulties that go beyond word access, word retrieval and the phonological representations of words, pointing to suboptimal semantic representations. Little is known about the underlying cause of lexical-semantic deficits in children with DLD. Often put forward as a possible cause is poor phonological short-term memory, which is considered an important prerequisite for vocabulary acquisition (Melby-Lervåg et al., 2012). There is extensive evidence of deficits in phonological short-term memory and verbal working memory in children with DLD (for a review, see Montgomery et al., 2010). Phonological short-term memory is often measured using a non-word repetition (NWR) task. Studies show that performance on NWR tasks correlates with word-learning skills in TD children (Gathercole et al., 1997) and in children with DLD (Alt & Plante, 2006).

The causal direction of the relationship between phonological short-term memory and word learning is not clear. Difficulties with phonological processing might lead

to poor phonological representations of words, which in turn may have a negative influence on the building of strong semantic representations. Indeed, NWR ability predicts vocabulary in young children between 4 and 5 years, but this relationship gets weaker in older children between 6 and 8 years (Gathercole et al., 1992; Gathercole, 2006). Furthermore, it has been found that vocabulary size is an important predictor of NWR ability, which could be explained as follows: as vocabulary size grows, phonological representations strengthen, which would improve non-word repetition ability (Metsala, 1999). Other studies fail to find evidence for a causal relationship between NWR ability and vocabulary. For example, Melby-Lervåg et al. (2012) carried out a large longitudinal study and did not find evidence for a causal relationship between NWR skills and vocabulary development in 4 to 7-year-old children. The authors also re-analyzed data from a similar longitudinal study (Gathercole et al., 1992), and failed to find the causal relationship that the authors of the original study had claimed. Finally, intervention studies have failed to find an effect of phonological memory-training on vocabulary knowledge (Melby-Lervåg et al., 2012: Dahlin et al., 2008, Schmiedek et al., 2010). Thus, although the difficulties in phonological processing in DLD are well-established, the role they play in vocabulary development remains unclear.

Statistical learning and the development of the lexicon

To summarize, a large body of studies points towards an important role for statistical learning in the acquisition of language. In children with DLD, the ability of extracting regularities from input seems to be affected, which could explain their language deficits. In this section we discuss the relationship between statistical learning and the development of the lexicon. Specifically, we look at the link between statistical learning and lexical-semantic knowledge.

Children with better statistical learning skills often have a larger vocabulary (Spencer et al., 2015), and Shafto et al. (2012) and Ellis et al. (2014) report a predictive relationship between TD infants' performance on a visual statistical learning task and their vocabulary size at a later point in time. In another longitudinal infant study, Singh et al. (2012) found that statistical learning ability in a word segmentation task at 7 months predicts productive vocabulary at 24 months.

Evidence also suggests a relationship between statistical learning and vocabulary in children with DLD. Evans et al. (2009) reported a correlation between statistical learning ability and vocabulary knowledge and claimed that lexical impairments might be explained by statistical learning difficulties. In another study, Mainela-Arnold and

Evans (2014) report a significant correlation between statistical learning ability on a word segmentation task and performance on a lexical-phonological access task. During this forward gating task, children heard increasingly longer parts of a word and had to guess which word they heard. On the other hand, no evidence was found for a relationship between statistical learning and performance on a word definition task. The authors suggest (from a comparison of their two p-values) that statistical learning underlies the acquisition of sequential lexical-phonological knowledge, but that lexical-semantic abilities might depend on other learning/memory systems.

The link between statistical learning and lexical-semantic knowledge requires further investigation. In the study of Mainela-Arnold and Evans (2014), the status of a potential relation cannot be concluded from comparing a null result with a statistically significant result. Moreover, they used a word definition task to measure lexical-semantic knowledge, which requires very explicit semantic knowledge. It could be the case that statistical learning is related to more implicit forms of semantic knowledge. Furthermore, statistical learning in this and many other studies was measured using a word segmentation task. It is not unexpected that this type of sequential statistical learning contributes to lexical-phonological knowledge due to the nature of the task. However, as Mainela-Arnold and Evans (2014) also state, it is possible that other types of (non-sequential) statistical learning that were not taken into account play a role in the building of a semantically rich lexicon.

Statistical learning mechanisms indeed seem to be sensitive to semantic information (see Paciorek & Williams (2015) for a review). For example, the mapping of newly learned words to their corresponding referents is suggested to be a gradual statistical learning process named cross-situational learning, which entails the (implicit) tracking of co-occurrences between words and their visual referents (Kachergis et al., 2014; Smith & Yu, 2008; Suanda et al., 2014; Yu & Smith, 2011). In another strand of research, Goujon (2011) showed that adults implicitly learn that the semantic categories of real-world scenes predict the position of the following target in a visual search task, indicating that semantic information is processed automatically and can be facilitated to make unrelated decisions. Similarly, Rogers et al. (2020) report that higher-order categories influence the learning of visual statistical regularities: people learn implicit mappings between visual stimuli better when the stimuli belonged to the same category rather than two different categories.

An important phenomenon in the development of the lexicon is shape bias. This entails the tendency for children to extend the use of newly learned object names to objects that share the same shape with the original object rather than the same color or size. The emergence of this shape bias might depend on statistical learning mechanisms: if children pick up the regularity that early learned object categories often share the same shape, they learn to consider shape as an important cue when learning new object labels. Results from a novel object name learning experiment of Collisson

et al. (2015) indicate that 3-to-4-year-old children with DLD do not show shape bias to a similar extent as TD children. Moreover, children with DLD perform more poorly on a task that measures visual paired-associate learning, and this performance predicts the strength of their shape bias. This finding suggests that an impairment in visual statistical learning might underlie the lagging development of shape bias in these children, which in turn may hinder their lexical development.

Another process in the development of the lexicon that could be supported by statistical learning mechanisms is learning to categorize and name the enormous number of different objects in the visual world. For example, a child needs to learn which round fruits are called apples and which ones are called peaches. Studies point out that infants automatically track the co-occurrence of visual features of objects in visual statistical learning tasks (Wu et al., 2011, 2010). This ability of learning which object features co-occur and which do not, plays an important role in learning about visual categories (Palmeri & Gauthier, 2004). Similarly, Younger (1985) and Plunkett et al. (2008) showed that statistical learning may underlie semantic category learning, as infants learn object categories based on the co-occurrence of features.

Distributional learning

A specific type of non-sequential statistical learning, distributional learning, plays a role in the formation of new categories as well. Maye et al. (2002) showed that infants can pick up speech sound categories based on the frequency distribution of speech sound exemplars. Their infants were exposed to variants from the /ta/-/da/ continuum. The distribution of the variants was either bimodal or unimodal: in the bimodal condition there were two distributional peaks, reflecting two distinct sound categories /t/ and /d/, while in the unimodal condition there was only one peak reflecting one broad category. After familiarization it was tested whether the infants could discriminate the endpoint tokens of the continuum. Maye et al. found that only their participants in the bimodal condition had statistically significantly formed two distinct categories, as they were able to discriminate the two endpoint tokens, while infants in the unimodal condition did not reach significance. This result indicated to Maye et al. that infants can learn phonetic categories based on distributional information. Although Maye et al.'s claim was based on a p-value comparison (a direct comparison between the two groups gave a non-significant p-value of 0.063), together with later findings of distributional learning of sound categories (Escudero et al., 2011; Hayes-Harb, 2007; Maye et al., 2008; Vandermosten et al., 2019; Wanrooij et al., 2014) the results point towards a distributional learning mechanism underlying bottom-up categorization of speech sounds.

More recent studies have shown that distributional learning mechanisms also play a role in the visual domain, for example in categorizing new faces. In the study of Alt-

vater-Mackensen et al. (2017), infants were subjected to a familiarization phase in either a unimodal or a bimodal condition. They saw tokens from a continuum that was created from two female faces. After familiarization, results from a discrimination task indicated that infants in the bimodal condition form two distinct categories of faces, while infants in a unimodal condition form one broad category. The same result has been shown in a novel visual object category learning experiment (Junge et al., 2018): infants in the bimodal condition showed better discrimination of two endpoint tokens than infants in the unimodal condition. Distributional learning thus seems to be important for the categorization of different types of sensory stimuli: speech sounds, faces and novel objects. To our knowledge, children with DLD have never been tested on such distributional learning tasks. In the current study we aim to investigate whether these children have a deficit in visual distributional learning and whether this ability correlates with their lexical-semantic knowledge, as a lessened sensitivity to regularities in object categories could contribute to their problems in building strong semantic representations.

The current study

Children with DLD have previously displayed difficulties with verbal and visual statistical learning which could hinder their ability to pick up language efficiently. Indeed, statistical learning ability correlates with or even predicts different types of linguistic skills, such as lexical skills. However, the relationship between statistical learning and the development of vocabulary skills in children with and without DLD is not well understood. In the current study we want to explore this relationship further by investigating visual distributional learning and its relation to vocabulary in children with and without DLD.

Our first research question was: are children with DLD less sensitive to distributional cues compared to TD children when learning novel visual object categories in an experiment? Distributional learning has never been investigated in individuals with DLD, but one study shows that distributional learning of speech sounds is impaired in children with dyslexia (Vandermosten et al., 2019). Developmental dyslexia and DLD are distinct but overlapping disorders (Snowling et al., 2020) and together with previous evidence showing that both verbal and visual statistical learning is impaired in children with DLD, we expected that they show less proficiency in visual distributional learning as well.

Our second research question was: Does the ability of visual distributional learning contribute to lexical knowledge in children with DLD? The underlying cause of the lexical-semantic difficulties in this group is not clear. There is extensive evidence for problems with phonological short-term memory, but this does not seem to be an adequate explanation. We expected that visual distributional learning contributes to

these lexical-semantic difficulties, as it could be important for learning semantic information about (the use of) words, object categories and how to map words to objects. Difficulties with processing visual patterns in the environment might result in problems with building a semantically rich lexicon.

To answer our research questions we constructed a visual distributional learning task based on Junge et al. (2018) to test novel object categorization in children with and without DLD. Moreover, we measured lexical knowledge comprehensively in the children with DLD: besides productive and receptive vocabulary size, we tapped the organization of the lexicon and the knowledge of relationships between concepts/words. Finally, we control for variation in phonological processing, as children with DLD are known to have difficulties with this ability and because it is probably related to lexical knowledge. We also controlled for variation in non-verbal intelligence.

Wanrooij et al. (2015) discuss potential pitfalls in the typical design employed when comparing a unimodal with a bimodal familiarization phase in distributional learning tasks. We therefore adapted a different design. In the usual design there might be a confounding factor at play: besides the number of distributional peaks in the input, the spreading of variants (or *dispersion*) also differs between conditions. This difference might result in easier discrimination of endpoint tokens for individuals who had been familiarized with the bimodal condition, as spreading of the variants is higher in that condition. Chládkova et al. (2020) designed a (auditory) distributional learning task that tackled this problem: they constructed two bimodal learning conditions which differed in the position of the distributional peaks, ensuring that spreading of the variants was not different in the two conditions. We applied this design to the visual distributional learning task of Junge et al. (2018).

Method

Participants

27 children diagnosed with DLD participated in our research. One child did not finish the statistical learning task and another child was removed because of bilingualism, resulting in a final sample of 25 children with DLD (17 male, 8 female) between the ages of 7;2 and 9;3 (years;months). For the control group we used previously collected data from a study in which TD children were tested on the same task (Broedelet et al., 2021).[1] We selected 25 children (15 male, 10 female) from a larger sample that

[1] We had planned to test a new group of TD children matched to the DLD group. Unfortunately, we were unable to administer the tests as all primary schools in the Netherlands were closed from March

matched the DLD group best regarding age and gender. Their ages varied between 7;6 and 8;9. Age did not differ significantly between groups (TD age in months $M = 97.64$, $SD = 4.99$, DLD age in months $M = 96.56$, $SD = 6.49$), as tested with a two-sample t-test: $t = 1.864$, $p = 0.063$.

The children with DLD were recruited via different institutions in the Netherlands: Pento, Royal Dutch Auris Group and VierTaal. All children had been officially diagnosed with DLD by a professional clinician and were included if they met the standard DLD inclusion and exclusion criteria used within the institution. All children met the following criteria: they scored at least 1.5 standard deviations below the age norm on at least two of the four language domains (speech, auditory processing, grammar, lexical-semantic development), tested with standardized tests like the CELF; their language disorder was not secondary to a physiological or neurological disorder such as ASD, ADHD or hearing difficulties; they did not have a severe form of dyspraxia and at least one of their caretakers had acquired Dutch as a native language. Data from one child was removed because he was growing up bilingually and answered multiple questions on a vocabulary task in English.

The TD children were recruited via two primary schools in the Netherlands and met the following criteria: they had not been diagnosed with hearing difficulties, language disorders, dyslexia, ADHD or ASD and had at least one caretaker that was a native speaker of Dutch. Our study was approved by the Ethical Committee of the Faculty of Humanities of the University of Amsterdam. The parents/caretakers of all children filled in an informed consent form prior to their participation.

To get a general estimate of the language ability in our DLD subgroup, we administered the Sentence Recalling subtask from the CELF (Clinical Evaluation of Language Fundamentals: Core Language Scales, Dutch version; Semel et al., 2010). In this task, children are asked to repeat sentences of increasing complexity, measuring their morphosyntactic abilities. The Raven Progressive Matrices task was administered to measure non-verbal intelligence (Raven, et al., 2003). One of the children could not finish the Sentence Recalling task due to time constraints. The children's scores (raw, percentile and if available norm and age-equivalent scores) on these two tasks are

to June 2020 due to the outbreak of COVID-19. After the reopening of the schools many restrictions still applied, making it impossible to enter schools for testing participants. We therefore decided to use a subset of an already collected dataset as control data. This dataset was previously used for an article about visual distributional learning in TD children (Broedelet et al., 2021). The decision to use previously collected data was taken only because of this circumstance, and *not* because we found a significant effect in this group and deemed it sufficient to use this data. As a result of this reuse, the control group, unlike the DLD group, was not tested on the background tasks measuring vocabulary, morphosyntactic skills, phonological processing and non-verbal intelligence. This means the control group could unfortunately not be matched on vocabulary skills to the DLD group.

shown in Table 1. The children with DLD had low scores on the Sentence Recalling task and performed on average 50 months below their age level, confirming that our sample indeed had difficulty with language acquisition, while they scored within the average range on non-verbal intelligence. This discrepancy between language skills and non-verbal cognitive skills is typical for children with DLD.

Table 1 – Scores of the children with DLD on the sentence recalling and non-verbal intelligence task.

Task	Raw scores	Norm scores	Percentile scores	AES	Diff.
Sentence Recalling (N=24)	4 .. 42 *M* = 18.46 *SD* = 9.27	1 .. 8 *M* = 3.58 *SD* = 2.02	0.1 .. 25 *M* = 4.07 *SD* = 6.37	36 .. 83 *M* = 45.79 *SD* = 13.05	-68 .. -21 *M* = -50.46 *SD* = 14.43
Raven's progressive Matrices	11 .. 38 *M* = 23.24 *SD* = 7.41		5 .. 95 *M* = 41.04 *SD* = 26.25		

Notes: AES = Age-equivalent score (months). Diff. = Difference AES and chronological age. The chronological age (months) is subtracted from the age equivalent score (months). A negative value means that the age-equivalent score was lower than the actual age (M = 96.56, SD = 6.61, range 86 - 111). Scale used for interpreting percentile scores: 0-3 Very low, 3-10 Low, 10-16 Below average, 16-84 Average, 84-90 Above average, 90-98 High, 98-100 Very high. The Sentence Recalling percentile score is in the low range; the Raven's percentile score is in the average range.

Stimuli and design distributional learning task

The design of this experiment follows Junge et al. (2018) and Chládkova et al. (2020), and was previously reported in Broedelet et al. (2021). The aim of our experiment was to measure whether the frequency distribution of tokens along a continuum influenced categorization of those tokens. To this end we constructed an 11-step continuum by morphing two pictures in equal steps using the Sqirlz 2.1 software (Xiberpic.com). We obtained permission to use the pictures of two cuddly toys from Giant Microbes (www.giantmicrobes.com) that were also used in the study of Junge et al. (2018). See Figure 1.

Figure 1 - Novel object continuum used in the experiment.

In the familiarization phase of the experiment, stimuli from the continuum were presented to the children. Two different between-participant familiarization conditions were constructed (see Figure 2). Both conditions contained a bimodal distribution, but the conditions differed concerning the position of the peaks in the continuum. Three of the 11 tokens, which were all equally frequent in both conditions, were used to measure categorization in the test phase: 6, 4 and 8, hereafter referred to as S (standard), D1 (deviant 1) and D2 (deviant 2).

In Condition 1 (Figure 2, blue line), token S and token D2 belonged to the same peak, while token 5 was shown less frequently, creating the perception of a category boundary. In Condition 2 (Figure 2, orange line), token S and token D1 belonged to the same peak and token 7 was shown less frequently. Our hypothesis was that our participants would learn that tokens in one distributional peak belong to one category while tokens from different peaks belong to two different categories. Therefore we predicted that children in Condition 1 learn that tokens S and D2 belong to one category while children in Condition 2 learn that tokens S and D1 belong to one category.

Children were shown 12 blocks of 24 stimuli each (288 stimuli in total), as well as 2 filler stimuli per block (see Figure 4). In each block, the tokens of the continuum were presented one by one following the frequency distribution shown in Figure 2, in a randomized order. Each stimulus was shown for 800 ms and the interstimulus interval was 200 ms (based on the results of Turk-Browne et al. (2005) and Arciuli & Simpson (2011)). Stimuli were shown against a gray background (see Figure 3). A cover task was added to the task to make it more engaging: the filler stimuli jumped across the screen and children were instructed to click on them as fast as possible.

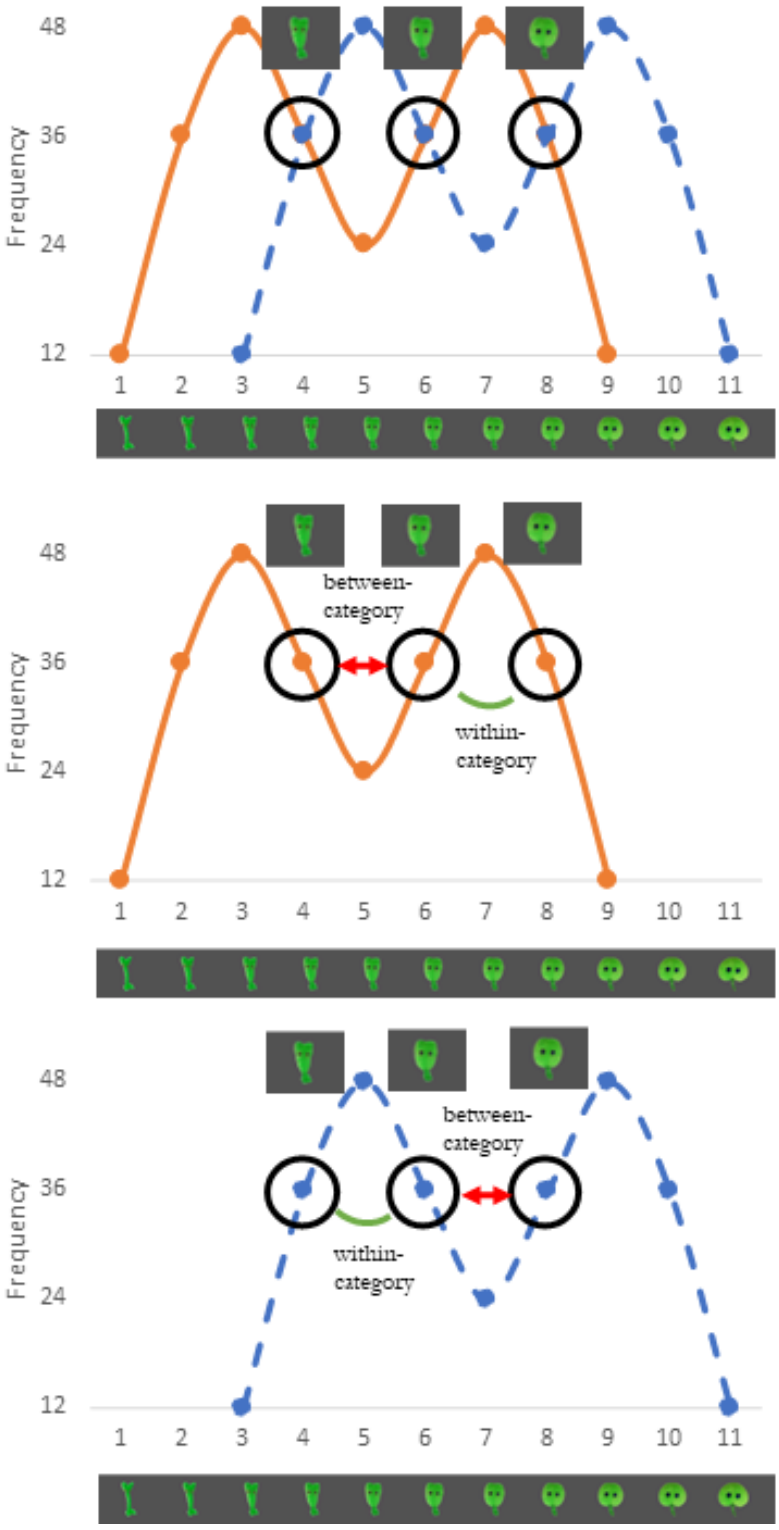

48
36
24
12
Frequency
1 2 3 4 5 6 7 8 9 10 11
between-category
within-category
between-category
within-category

Figure 2 - Familiarization conditions in the experiment. In Condition 1 (blue line), tokens S and D2 belong to one distributional peak while D1 lies in another peak. On the other hand, in Condition 2 (orange line), tokens S and D1 belong to one distributional peak while D2 lies in another peak. We hypothesize that participants in Condition 1 will learn that S and D2 belong to one category and thus will look more alike than S and D1, and the reversed for participants in Condition 2.

Figure 3 – A familiarization trial.

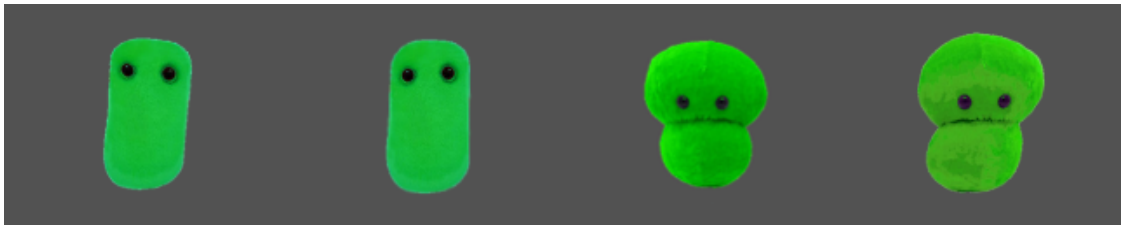


Figure 4 – Stimuli that were used as fillers/cover task.

Categorization was tested after familiarization using AXB-type questions. Children were asked to choose whether stimulus D1 or D2 looked more like stimulus S. In the

eight questions, stimulus S was shown above a white stripe and stimuli D1 and D2 were shown below the stripe (see Figure 5). The position of D1 and D2 (left/right) was counterbalanced. Four filler questions were included to add some variation to the test phase, as well as a practice question. For these questions the stimuli that functioned as fillers in the familiarization phase were used and there was a clearly correct answer. The test phase was identical for every child, except that the order of the test questions was randomized. We hypothesized that children that underwent Condition 1 of the familiarization phase would choose stimulus D2 more often than children in Condition 2. This effect of Condition would be considered a learning effect.

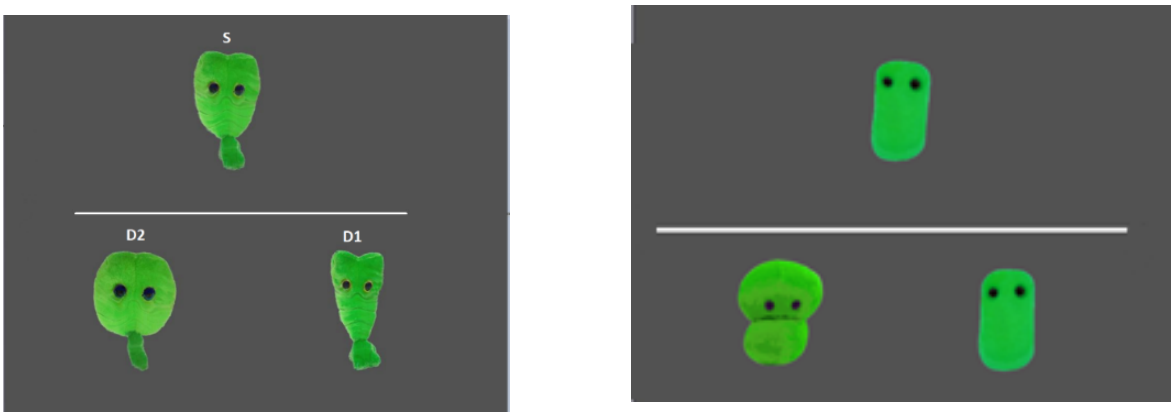


Figure 5 – A test question and filler/practice question.

Measures of vocabulary, phonological processing, non-verbal intelligence and socio-economic status

To investigate the relationship between visual distributional learning and lexical skills in children with DLD[2], we administered several subtests of the CELF (Active Vocabulary, Word Classes 1 or 2 (depending on the age of the child) and Word Associations, as well as the Peabody Picture Vocabulary Task (PPVT; Schlichting, 2005). The tasks were used as measures of receptive and productive vocabulary size (PPVT, Active Vocabulary), the ability to find and express semantic relations between words/concepts (Word Classes 1 and 2) as well as the ability to name words of a semantic category as an indicator of lexical-semantic organization (Word Associations). See Table 2 for more information about the vocabulary tasks.

As control tasks, the children were tested on phonological short-term memory using the digit span task Number Repetition 1 from the CELF, on verbal working memory

[2] Our original plan was to investigate this relationship in both groups of children. Unfortunately, as is mentioned in our first footnote, we were not able to test the TD children on these tasks.

using the Number Repetition 2 task (digit span backwards) from the CELF and on verbal short-term memory using the non-word repetition task (Rispens & Baker, 2012). Moreover, performance on the Raven Progressive Matrices task was used as a control variable for non-verbal intelligence. See Table 3 for more information about the control tasks. Finally, as socio-economic status (SES) may play a role in vocabulary development (e.g. Hoff, 2003), we took the SES of the children into account using a database from Sociaal en Cultureel Planbureau (2018). In this database, socio-economic scores are computed on the basis of the average education level and income in a particular zip code. The SES scores are based on the home addresses of the children.

Table 2 – Vocabulary measures administered to the children with DLD.

Construct	Task	Description	Scoring	Score range
Vocabulary size	Receptive vocabulary (PPVT)	Children heard a word and had to point to one of the four pictures.	Correct: 1 point Incorrect: 0 points	0 .. 204
Vocabulary size	Productive vocabulary (CELF)	Children saw a picture and had to name it.	2 points for a correct answer, for some items there were 1-point answer possibilities	0 .. 56
Semantic knowledge	Word Classes 1 (7 y.o. children) (CELF)	Children had to choose which two out of three/four pictures were related and why.	1 point for choosing the correct picture, 1 point for expressing the relationship correctly	0 .. 38
Semantic knowledge	Word Classes 2 (8+) (CELF)	Children had to choose which two words out of four were related and why.	1 point for choosing the correct word, 1 point for expressing the relationship correctly	0 .. 40
Lexical-semantic organization	Word Associations (CELF)	Children had to name as many words as they could in a semantic category: food, clothes and professions.	1 point for every related word	0 .. ∞

Table 3 – Control measures administered to the children with DLD.

Construct	Task	Description	Scoring	Score range
Verbal short-term memory	Digit span forwards	Children had to repeat strings of number increasing in length.	Correct: 1 point Incorrect: 0 points	0 .. 16
Verbal working memory	Digit span backwards	Children had to repeat strings of number backwards increasing in length.	Correct: 1 point Incorrect: 0 points	0 .. 14
Phonological short-term memory	Non-word repetition	Children had to repeat non-words.	Correct: 1 point Incorrect: 0 points	0 .. 22
Non-verbal intelligence	Raven Progressive Matrices	Children had to complete a visual pattern.	Correct: 1 point Incorrect: 0 points	0 .. 60

Procedure

Testing took place in a quiet room in the school or in the home of the child. The distributional learning experiment was run on a laptop computer using E-Prime 3.0 (Psychology Software Tools, Pittsburgh, PA). Children wore headphones. We had recorded the instructions in advance, in a child-directed manner. Before the experiment started, the children were instructed to look at the images on the screen and click on moving images as fast as they could if they saw one. They were told to watch carefully as there would be questions about the images later on, but the type of questions was not specified. The experiment started when the child confirmed that s/he understood the task. Familiarization condition was counterbalanced between participants. There was a short break halfway the familiarization phase and the child could indicate when s/he wanted to continue. The test phase started immediately after the familiarization phase with a practice question. Children were instructed to carefully look at the image above the white stripe, and to indicate which of two images below the stripe they thought looked more like the upper image. The experimenter repeated the question while pointing out the images. The experiment had a total duration of approximately 10 minutes.

Besides the distributional learning task, the children with DLD did two other statistical learning tasks (results are not discussed in this paper) as well as the aforementioned background tasks. For those children, testing was divided over two separate test sessions on different days; the second session usually took place within a few days

or one week. The order of the tasks within the sessions as well as the order of the sessions was counterbalanced across participants. Each test session took approximately 50 to 60 minutes.

Results

Split-half reliability distributional learning task

Split-half reliability was computed as a measure of reliability of the distributional learning task. Two separate generalized mixed effect models were run with only the odd or even test items included. Then, the correlation between the answers to even and odd test items was computed, using the random slopes of the intercept for the even/odd test items. After the application of the Spearman-Brown correction, the split-half reliability of the task turned out to be $r = 0.73$ (95% CI 0.52 .. 0.85), approaching the value of $r = 0.80$ which is considered the standard that reliable tests should meet (Nunnally & Bernstein, 1994).

Group comparison distributional learning task[3]

See Table 4 and Figure 6 for the descriptive data. As a first step in our analysis, we removed all practice and filler items from the data. A generalized mixed effect model was run with the package *lme4* (Bates et al., 2015) in R (R Core Team, 2020) to test whether familiarization condition and participant group influenced categorization. The choice for stimulus D2 (which could either be 1 or 0) was the dependent variable. Between-participant predictors were Condition (Condition 1 or 2), Group (TD or DLD) and Age (in months). PositionD2 was a within-participant predictor reflecting the position of token D2 (left or right) that varied between test items. We chose the maximal model that is still correctly computable and that keeps all its included predictors and interactions reportable (by including random slopes for all within-participant predictors and interactions). The model includes main effects for Condition, Group, Age and PositionD2, all two- and three-way interactions between Condition, Group and Age as well as the simple interaction between Condition and PositionD2. Moreover, we included random intercepts by participant as well as by-subject random slopes for PositionD2. Sum-to-zero orthogonal coding (Kraemer & Blasey, 2004) was applied to the predictors Condition (-½ for Condition 2 and +½ for Condition 1), Group (-½ for DLD and +½ for TD) and Position D2 (-½ for right and +½ for left). The predictor Age was centered by subtracting its average.

We predicted that if children are sensitive to the distributional cues in the familiarization phase, our children in Condition 1 would prefer the combination S + D2, while

[3] The TD children of whom results are reported here are a subgroup of the sample reported in Broedelet et al. (2021).

our children in Condition 2 would prefer the combination S + D1 - in other words, a stronger preference for D2 in Condition 1 than Condition 2. This could manifest as a significant effect of Condition on the dependent variable. Moreover, we expected that our children with DLD would be less sensitive to the distributional cues in the familiarization phase than our TD children, which could manifest as a significant interaction between the effects of Condition and Group on the dependent variable, indicating that the Condition effect is not equally strong in the two subpopulations.

Confirmatory results

In our sample, as determined by our model, Condition influenced the choice for stimulus D2: children in Condition 1 were 4.04 times more likely to choose stimulus D2 than children in Condition 2, and this effect was significantly above 1: $z = 2.758$, $p = 0.006$, 95% CI 1.497 .. 10.9. This is in line with our prediction and indicates that school-aged children can learn novel visual object categories based on distributional properties. Our second prediction is not confirmed: although the effect of Condition was 1.007 times stronger in the TD group compared to the DLD group, this interaction between Condition and Group was not significantly above 1: $z = 0.007$, $p = 0.994$, 95% CI 0.15 .. 6.8. We thus cannot conclude anything about a difference in distributional learning in children with DLD compared to TD children: the confidence interval tells us that children with DLD could be up to 6.7 times better or 6.8 times weaker on the visual distributional learning task than TD children. We therefore cannot conclude whether children with DLD do or do not have a distributional learning deficit.

Exploratory results

To explore whether children with DLD show a distributional learning effect, we ran a separate model which only included the children with DLD. This model included the main effects for Condition, Age and PositionD2 as well as all three-way interactions between those predictors. According to the model, our children with DLD in Condition 1 were 3.75 times more likely to choose D2 than our children in Condition 2, but the effect was not significantly above 1: $z = 1.788$, $p = 0.074$, 95% CI 0.86 .. 19.4[4]. On the basis of this result we cannot conclude whether children with DLD are able to learn novel visual object categories based on distributional information[5].

[4] When we ran a model which included random slopes per participant for PositionD2 (as we did in our first model with all participants), the effect of Condition was 4.11 (95% CI 1.01 .. 16.7): $z = 1.977$, $p = 0.048$. However, as this model had a singular fit, we chose to report the results of a simplified model without random slopes for PositionD2 (this makes the effect of PositionD2 unreportable, but as we are not directly interested in this effect, this is not problematic). Note that neither the p-value of 0.074 neither the p-value of 0.048 can be called statistically significant, because this exploratory test came on top of the earlier confirmatory test, for which we already used a preset p-value criterion of 0.05.

[5] We also ran an analysis that only included the TD children, which yielded a significant effect of Condition ($z = 2.047$, $p = 0.04$). However, please note that this finding cannot be interpreted as a difference

Table 4 - Descriptive data for the choice of stimulus D1 or D2.

	TD children		Children with DLD	
	D1	D2	D1	D2
Condition 1	55	49 <u>target</u>	61	35 <u>target</u>
Condition 2	71 <u>target</u>	25	84 <u>target</u>	20

in distributional learning between children with DLD and TD children, as the effect of Group was not significant in our first model.

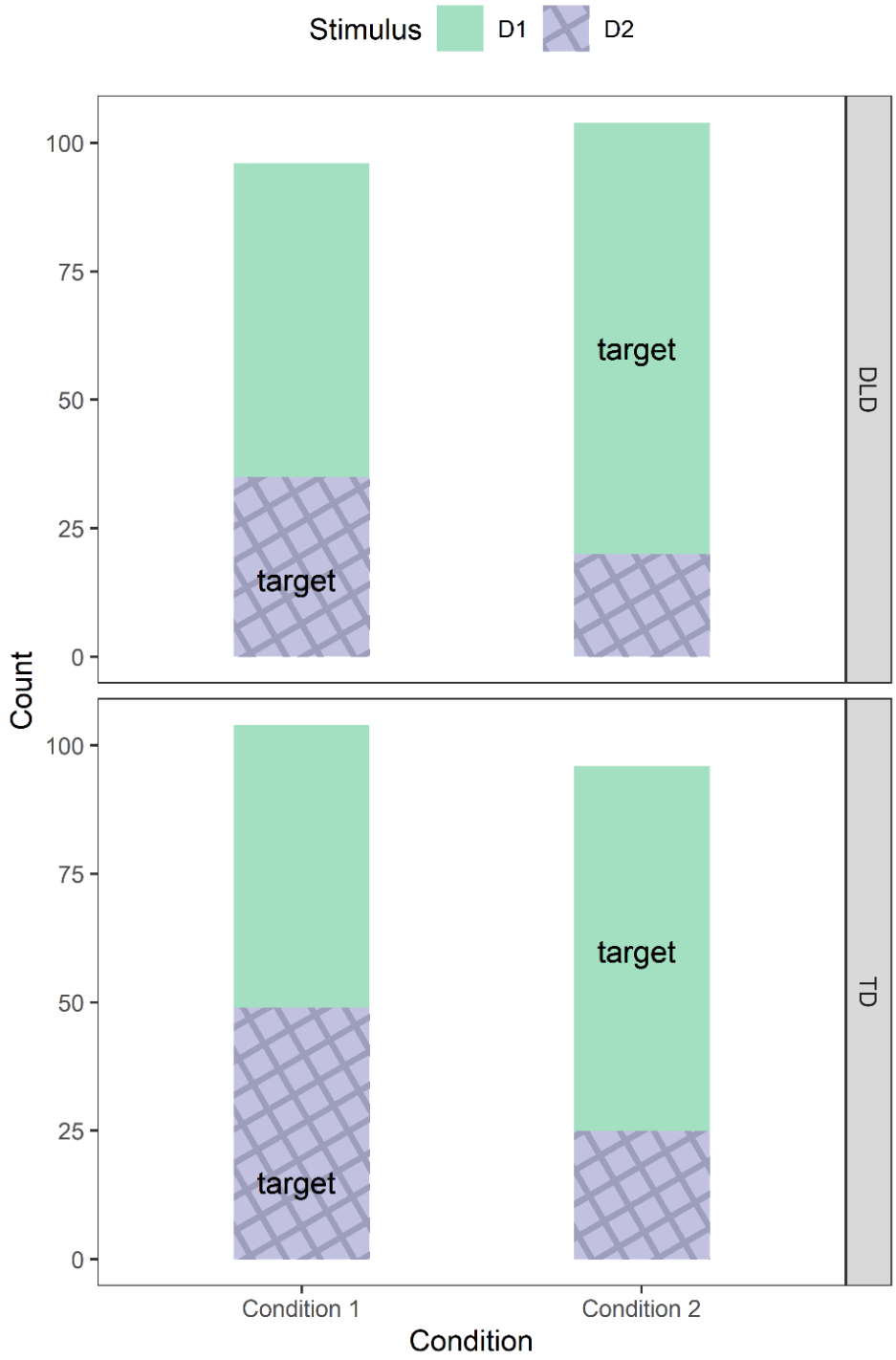


Figure 6 - Choice for stimulus D1 or D2 depending on condition and group.

Regression analyses

Descriptive data

To investigate the relationship between distributional learning and vocabulary, we administered tasks measuring several types of lexical knowledge to the children with DLD, as well as several control tasks (see section 4.3). In Table 5 we present the scores of the children with DLD on the vocabulary tasks and in Table 6 their scores on the control tasks: the raw scores, the norm and percentile scores (if available), and the age-equivalent scores. The raw scores are used in our statistical analysis. The norm, percentile and age-equivalent scores are presented to illustrate the abilities of the children with DLD.

Table 5 – Children with DLD's scores on the vocabulary task.

Task	Subtask	Raw scores	Norm scores	Percen-tiles	AES	Diff.
Productive vocabulary		8 .. 41 *M* = 28.16 *SD* = 8.94	2 .. 12 *M* = 6.84 *SD* = 2.46	0.4 .. 75 *M* = 20.46 *SD* = 21.33	36 .. 98 *M* = 73.24 *SD* = 16.69	-62 .. 7 *M* = -23.32 *SD* = 15.83
Receptive vocabulary		70 .. 119 *M* = 90.48 *SD* = 13		0 .. 91 *M* = 27.36 *SD* = 26.82		
Word asso-ciations		10 .. 42 *M* = 23.92 *SD* = 6.37	2 .. 15 *M* = 7.48 *SD* = 2.45	0.4 .. 95 *M* = 24.22 *SD* = 19.77	42 .. 133 *M* = 77.2 *SD* = 18.42	-56 .. 42 *M* = -19.36 *SD* = 19.19
Word clas-ses	Recep-tive	2 .. 19 *M* = 11.2 *SD* = 6.95	3 .. 12 *M* = 7.24 *SD* = 2.63	1 .. 75 *M* = 24.92 *SD* = 22.6	36 .. 109 *M* = 70.36 *SD* = 22.85	-68 .. 18 *M* = -26.2 *SD* = 25.59
	Expres-sive	0 .. 18 *M* = 8.8 *SD* = 5.95	1 .. 13 *M* = 6.88 *SD* = 2.71	0.1 .. 84 *M* = 21.68 *SD* = 22.84	36 .. 116 *M* = 71.92 *SD* = 19.19	-68 .. 25 *M* = -24.64 *SD* = 21.06
	Total	2 .. 37 *M* = 20 *SD* = 12.79	2 .. 13 *M* = 6.88 *SD* = 2.60	0 .. 84 *M* = 21.3 *SD* = 22.57	36 .. 116 *M* = 71.6 *SD* = 19.4	-68 .. 25 *M* = -24.96 *SD* = 21.72

Notes: AES = Age-equivalent score (months). Diff. = Difference AES and chronological age. The chronological age (months) is subtracted from the age equivalent score (months). A negative value means that the age-equivalent score was lower than the actual age (M = 96.56, SD = 6.61, range 86 - 111). Scale used for interpreting percentile scores: 0-3 Very low, 3-10 Low, 10-16 Below average, 16-84 Average, 84-90 Above average, 90-98 High, 98-100 Very high. The scores for the vocabulary tasks fall within the average range.

Table 6 – Children with DLD's scores on the control tasks

Task	Subtask	Raw scores	Norm scores	Percentile scores	AES	Diff.
Raven's progressive Matrices		11 .. 38 *M* = 23.24 *SD* = 7.41		5 .. 95 *M* = 41.04 *SD* = 26.25		
Digit Span	Forwards	3 .. 9 *M* = 5.36 *SD* = 1.58	1 .. 12 *M* = 6 *SD* = 2.8	0.1 .. 75 *M* = 16.6 *SD* = 21.27	50 .. 103 *M* = 68.76 *SD* = 16.87	-52 .. 12 *M* = -27.8 *SD* = 17.33
	Backwards	0 .. 4 *M* = 2.72 *SD* = 1.02	2 .. 11 *M* = 7.52 *SD* = 2.35	0.4 .. 63 *M* = 26.06 *SD* = 19.34	57 .. 101 *M* = 79.52 *SD* = 13.97	-43 .. 14 *M* = -17.04 *SD* = 13.73
	Total	4 .. 12 *M* = 8.08 *SD* = 1.91	1 .. 10 *M* = 5.68 *SD* =2.39	0.1 .. 50 *M* = 12.9 *SD* = 14.67	48 .. 102 *M* = 71.8 *SD* = 11.81	-56 .. -2 *M* = -24.76 *SD* = 12.31
Non-word repetition		0 .. 9 *M* = 3.36 *SD* = 2.36		Low		

Notes: AES = Age-equivalent score (months). Diff. = Difference AES and chronological age. The chronological age (months) is subtracted from the age equivalent score (months). A negative value means that the age-equivalent score was lower than the actual age (M = 96.56, SD = 6.61, range 86 - 111). Scale used for interpreting percentile scores: 0-3 Very low, 3-10 Low, 10-16 Below average, 16-84 Average, 84-90 Above average, 90-98 High, 98-100 Very high. The scores for the Raven, and digit span backwards fall within the average range, the scores for digit span forwards and total digit span score fall in the below average range.

In contrast to their scores on the sentence recall task (see Table 1), the children with DLD scored within the average range (low end of the continuum) on the measures of vocabulary. The age-equivalent scores on these subtasks were between 19.36 and 26.2 months below their chronological age. Their non-verbal intelligence scores are also within the average range (see Table 5). However, the children showed below-average scores on the digit span forward task, which presumably reflect limitations in phonological short-term memory, which are reported often in DLD (Montgomery et al., 2010). Norm scores are available for the non-word repetition task for TD children of 7 (*N* = 96) years old, 8 years old (*N* = 82) and 9 years old (*N* = 208)[6]. The mean raw scores

[6] https://progracy.com/normscores/

for these age groups are 8.03, 8.83 and 9.07 out of 22 words correct respectively. Compared to that, the average score of 3.36 out of 22 in our group of children with DLD (see Table 6) can be considered as low. The children's age in months was on average 96.56 (*SD* = 6.61, range 86 .. 111), and their SES score on average -0.37 (*SD* = 1.04, range = -1.96 .. 1.52).

Principle component analysis

Prior to the regression analysis, all variables were centered around zero and scaled to a standard deviation of 1. To reduce the number of predictor variables, we ran a principal component analysis (PCA) in R using the raw scores on the digit span forward, digit span backward, non-word repetition and non-verbal intelligence tasks. The PCA analysis yielded four components, which explained 44%, 36%, 15% and 5% of the variance respectively. On the basis of this outcome, we decided to use three components, as they together explained 95% of the variance in the data. After varimax rotation, the three components explained 46%, 27% and 26% of the variance respectively. These components were saved and used for further analysis. See Table 7 for the component loadings. The first component represented phonological processing (mainly digit span forward and non-word repetition scores, the scores of which strongly correlated (r = 0.77, p = 0.0001)), the second component non-verbal intelligence (mainly Raven scores), and the third component verbal working memory (mainly digit span backward scores).

Table 7 – Standardized loadings of varimax-rotated PCA.

	Component 1 (phonological processing)	Component 2 (non-verbal intelligence)	Component 3 (verbal working memory)
Digit span forwards	0.93	-0.22	0.05
Digits span backwards	0.05	0.20	0.98
Non-word repetition	0.95	0.13	0.03
Non-verbal intelligence	-0.05	0.97	0.21

Predictor variables

The predictor variables were accuracy on the distributional learning task, age, SES, and the three component scores representing phonological processing, non-verbal intelligence and verbal working memory respectively. There were no significant correlations between the predictor variables (see Table 8). Accuracy on the distributional learning task was used as the measure for distributional learning ability, and was computed by comparing the answer to every test question to the target answer. For Condition 1, the target answer was D2 while it was D1 for Condition 2. This variable thus reflects sensitivity to the distributional properties in the familiarization phase. See Figure 7 for the distribution of the accuracy scores.

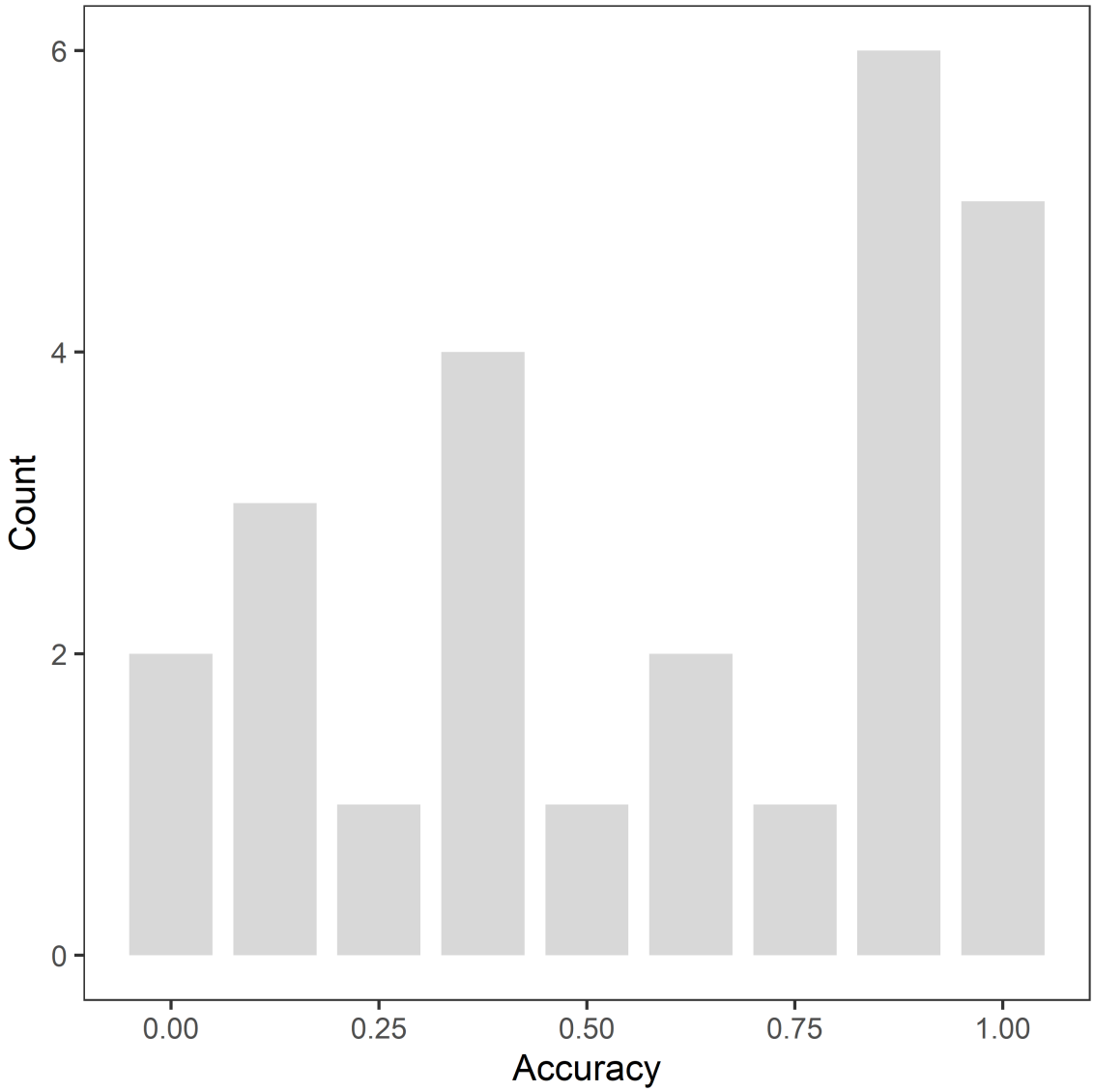


Figure 7 – Distribution of accuracy scores on the distributional learning task.

Table 8 - Correlations between the predictor variables.

	Comp 1 (phonological processing)	Comp 2 (non-verbal intelligence)	Comp 3 (verbal working memory)	Age	SES
Distributional learning	r = -0.17 p = 0.426	r = -0.05 p = 0.819	r = -0.24 p = 0.256	r = 0.09 p = 0.677	r = 0.03 p = 0.881
Component 1 (phonological processing)		r = 0 p = 1	r = 0 p = 1	r = 0.09 p = 0.675	r = 0.02 p = 0.917
Component 2 (non-verbal intelligence)			r = 0 p = 1	r = -0.21 p = 0.323	r = 0.02 p = 0.923
Component 3 (verbal working memory)				r = 0.26 p = 0.217	r = -0.22 p = 0.299
Age					r = 0.04 p = 0.866

Dependent variables

We ran four separate multiple linear regression analyses in R to test the relationship between distributional learning and different measures of vocabulary. The dependent measures were raw scores on the tasks measuring receptive vocabulary size, productive vocabulary size, and word associations. For the scores on the word classes tasks (part 1 and part 2) we decided to use the norm total scores (receptive + expressive) instead of raw scores (see Table 5)[7].

Regression analyses

The first model was run with receptive vocabulary size as the dependent variable and the five predictors as predictor variables. The model did not explain variation in receptive vocabulary size better than the null model (F = 0.59, p = 0.734) and none of the predictors were significant (see Table 9). The second model with productive vocabulary size as the dependent variable also was not significant (F = 1.693, p = 0.18) and contained no significant predictors (see Table 10). The third model with word classes total score as the dependent variable was not significant (F = 1.604, p = 0.2033), but component 2 (non-verbal intelligence) significantly predicted word classes score (t = 2.156, p = 0.045), indicating that the ability of completing non-verbal patterns might

[7] We felt it was not possible to use the raw scores, as the two parts of the task (part 1 for children up until 7 years old and part 2 for 8+ children) yielded different ranges of scores, while they were meant to measure the same underlying skill. Using the norm scores enabled us to use word category score as one variable for the whole group of children. The norm scores were computed from tables provided in the manual of the test, based on a sample of 1336 Dutch children (5-16 years old). The norm score provides information about a child's performance compared to the age norm.

explain unique variance in semantic knowledge about words, but please note that this result is exploratory. None of the other predictors were significant (see Table 11). The last model with word association score as the dependent variable was not significant ($F = 0.827$, $p = 0.564$), and none of the variables significantly predicted the dependent variable (see Table 12). In none of the models distributional learning significantly predicted vocabulary scores. Based on this null result, we cannot conclude anything about the relationship between visual distributional learning and vocabulary knowledge.

Table 9 – Results from the first linear model predicting receptive vocabulary size.

Predictor	Estimate (log odds) [95% CI]	Std. error (log odds)	*t*	*p*
Age	0.18 [-0.31 .. 0.67]	0.234	0.782	0.444
SES	0.09 [-0.37 – 0.56]	0.222	0.420	0.680
Component 1 (phonological processing)	0.27 [-0.20 .. 0.73]	0.220	1.204	0.244
Component 2 (non-verbal intelligence)	0.17 [-0.30 .. 0.63]	0.221	0.747	0.465
Component 3 (verbal working memory)	0.12 [-0.38 .. 0.62]	0.239	0.501	0.622
Distributional learning	0.15 [-0.33 .. 0.63]	0.229	0.674	0.509
Comparison with null model: $F = 0.59$, $p = 0.734$				

Table 10 – Results from the second linear model predicting productive vocabulary size.

Predictor	Estimate (log odds) [95% CI]	Std. error (log odds)	*t*	*p*
Age	0.32 [-0.11 .. 0.75]	0.205	1.584	0.131
SES	0.36 [-0.05 .. 0.77]	0.194	1.869	0.078
Component 1 (phonological processing)	0.29 [-0.11 .. 0.70]	0.193	1.520	0.146
Component 2 (non-verbal intelligence)	0.14 [-0.27 .. 0.55]	0.193	0.729	0.476
Component 3 (verbal working memory)	-0.03 [-0.47 .. 0.41]	0.209	-0.126	0.901

Distributional learning	0.08 [-0.34 .. 0.50]	0.200	0.401	0.693
Comparison with null model: $F = 1.693$, $p = 0.18$				

Table 11 – Results from the third linear model predicting word classes total score.

Predictor	Estimate (log odds) [95% CI]	Std. error (log odds)	*t*	*p*
Age	-0.19 [-0.63 .. 0.24]	0.207	-0.930	0.365
SES	0.04 [-0.38 .. 0.45]	0.196	0.180	0.8595
Component 1 (phonological processing)	-0.25 [-0.66 .. 0.16]	0.195	-1.301	0.2098
Component 2 (non-verbal intelligence)	0.42 [0.01 .. 0.83]	0.195	2.156	0.045*
Component 3 (verbal working memory)	0.03 [-0.42 .. 0.47]	0.211	0.124	0.903
Distributional learning	-0.17 [-0.60 .. 0.25]	0.202	-0.859	0.402
Comparison with null model: F = 1.604, p = 0.2033				

Table 12 – Results from the fourth linear model predicting word association score.

Predictor	Estimate (log odds) [95% CI]	Std. error (log odds)	*t*	*p*
Age	0.14 [-0.33 .. 0.62]	0.227	0.630	0.536
SES	0.23 [-0.23 .. 0.68]	0.215	1.049	0.308
Component 1 (phonological processing)	0.10 [-0.34 .. 0.55]	0.214	0.486	0.633
Component 2 (non-verbal intelligence)	0.13 [-0.32 .. 0.57]	0.214	0.584	0.567
Component 3 (verbal working memory)	-0.30 [-0.79 .. 0.18]	0.232	-1.307	0.208
Distributional learning	0.88 [-0.38 .. 0.55]	0.222	0.398	0.695
Comparison with null model: $F = 0.827$, $p = 0.564$				

Discussion

In the current study we aimed to shed more light on the relationship between statistical learning ability and lexical-semantic skills in children with and without DLD. Specifically, we investigated whether children with DLD are sensitive to distributional information in a visual distributional learning task, and whether this ability is related to different types of lexical knowledge. Our results show that, overall, school-aged children learn novel visual object categories based on distributional information. We cannot answer our first research question as we did not find evidence for or against a visual distributional learning deficit in children with DLD. The confidence interval of our group comparison shows that children with DLD could be between 6.8 times weaker and 6.7 times better on the visual distributional learning task than TD children. The finding of a non-significant group difference could be due to chance. It is possible that the true effect is zero, but we can only speculate about possible underlying reasons.

It could be the case that children with DLD have no disadvantage in visual distributional learning compared to TD children. Previous evidence has suggested that visuomotor statistical learning is impaired in children with DLD (Lum et al., 2014; Obeid et al., 2016; Tomblin et al., 2007). However, null results have also been found (Aguilar & Plante, 2014; Noonan, 2018) and Lammertink et al. (2020) report evidence for visual statistical learning in children with DLD. Intact visual statistical learning cannot be concluded from our null result, but accumulated evidence could point towards a specifically verbal statistical learning deficit in children with DLD, as opposed to a domain-general deficit. Statistical learning is often characterized as a domain-general ability, but research suggests the existence of different domain-specific components of statistical learning (Siegelman, 2020). It is also possible that sequential statistical learning as is tested with for example word segmentation tasks is problematic for children with DLD, while specifically distributional learning is not. More research is necessary to disentangle these possibilities. For example, it would be interesting to investigate whether *verbal* distributional learning is problematic for children with DLD.
The absence of a significant DLD–TD difference could also be due to a lack of statistical power. We tested 25 children in both participant groups, but the between-participants design of our experiment results in relatively limited number of participants per subgroup. Future studies should test larger participant groups and/or change the design such that multiple between-participant comparisons are avoided. Another option would be to test categorization in a way that would provide more data, for example by using an online behavioral measure or an neurological measure like EEG (Altvater-Mackensen et al., 2017), which could make the task more sensitive to potential DLD–TD differences.

To answer our second research question, we investigated whether distributional learning ability predicted vocabulary knowledge in children with DLD, while controlling for variation in phonological processing, verbal working memory, non-verbal intelligence, SES and age. We did not find any evidence for or against this relationship in our sample of children with DLD. Apart from chance, several factors could underlie this null-result. It could be the case that, as statistical learning tasks are designed to measure group-level performance, they are not suitable for measuring individual differences reliably and thus should not be used to predict differences in language outcome.(Arnon, 2019; Siegelman et al., 2017; Siegelman et al., 2017). For example, Arnon (2019) showed that three different statistical learning tasks had a low test-retest reliability and internal consistency in children, illustrating that they did not capture individual statistical learning ability reliably. This is a serious problem in the field of statistical learning research, as correlations between statistical learning ability and language proficiency might have been both overestimated and underestimated in previously reported studies (Siegelman, 2020). The split-half reliability of our visual distributional learning task was $r = 0.73$, approaching the standard of $r = 0.80$. This suggests that the test is a fairly reliable test of categorization. However, test-retest reliability should still be investigated to find out whether this task is able to capture individual differences reliably.

Another phenomenon that could occur when investigating individual differences in statistical learning is a large portion of the participants performing around chance level. Variation around chance level is not meaningful variation, which could result in the absence of significant correlations. However, this does not seem to be the case for our sample (see Figure 7). Another problem with this type of tasks might be that implicit knowledge that is built during familiarization does not transfer to the more explicit test questions in the test phase. Introducing more implicit and/or online measures of statistical learning could address this problem.

Importantly, although we did not compare the children with DLD to TD children on measures of vocabulary directly, it is striking that the percentile scores of the children with DLD in our sample are within the average range. Still, it is important to note that the ranges are wide and the children do fall behind same-aged peers if we consider the age-equivalent scores. The scores on the task measuring syntax and morphology do fall in the low range. This could mean that grammatical difficulties are more pronounced than vocabulary problems in our sample. Future studies could consider picking specific subgroups of children with DLD who have pronounced vocabulary problems to investigate the relationship between statistical learning and vocabulary development.

Although we cannot conclude this on basis of our results, there is also the possibility that there is no (strong) relationship between statistical learning and lexical-semantic knowledge. Perhaps statistical learning does contribute to more structural linguistic

knowledge such as rules and regularities, but deeper (semantic) knowledge is subject to other types of learning mechanisms, although research did point out that statistical learning mechanisms are sensitive to semantic information (Goujon, 2011; Paciorek & Williams, 2015). Possibly, deficits in other cognitive mechanisms such as attention, inhibition or verbal short-term memory play a role in the lexical-semantic difficulties that are observed in children with DLD (Alt & Plante, 2006; Mainela-Arnold & Evans, 2014). More research into these difficulties and their underlying mechanisms is necessary.

We included measures of phonological processing, verbal working memory and non-verbal intelligence in our regression models as control variables. Somewhat unexpectedly, we did not find evidence for a contribution of phonological processing or verbal working memory ability to different types of vocabulary knowledge in our sample of children with DLD. Similarly, Rispens and Baker (2012) found no evidence for a relationship between non-word repetition and vocabulary size in TD children and children with DLD, and the longitudinal study of Melby-Lervåg et al. (2012) yielded no evidence of a causal relationship between non-word repetition and vocabulary acquisition in 4-7 year old TD children. A meta-analysis could shed light on the relationship between phonological processing and vocabulary development in children with and without DLD. Moreover, we found an indication that non-verbal intelligence contributes to word category knowledge in children with DLD. This might be explained by similarities between the tasks: in the Word Category task, children had to choose which two out of three pictures/words were related (and why), while in the Raven progressive matrices task children had to complete visual patterns (see Table 2). Still, it is an interesting finding that non-verbal intelligence could explain variation in the verbal (semantic) domain, although we want to emphasize that this is an exploratory finding.

A shortcoming of the visual distributional learning task we have used is the finding that children overall prefer the combination S + D1, which is a result we have also reported in Broedelet et al. (2021). In that study, we tested 32 adults in an online experiment to explore *a priori* preferences for either S+D1 or S+D2. We wanted to investigate how participants who had not been exposed to a familiarization phase would answer questions similar to the test phase of our experiment. Results showed that participants chose D1 to look more like S 75% of the time, which was significantly higher than chance level. This result implies that D1 looks more like S for most participants, which is not an ideal starting point for testing the influence of distributional learning on categorization. This a priori preference might have diminished the distributional learning effect as well as a potential group difference in learning. However, our results show that despite this preference for the combination of S+D1, exposure to a familiarization phase in which S and D2 belonged to one distributional peak still caused participants to categorize S and D2 more often. Future studies might choose to use different stimuli when testing visual distributional learning and test beforehand whether participants show any unexpected preferences.

Conclusion and future directions

Our study shows that school-aged children can learn novel visual object categories based on distributional information. We did not find evidence for or against a visual distributional learning deficit in children with DLD. Future research could use our results for meta-analyses. Moreover, it would be interesting to investigate whether children with DLD have a domain-general deficit in statistical learning or solely a verbal statistical learning deficit, for example by a comparison between visual and verbal distributional learning. The relationship between statistical learning and lexical-semantic knowledge should be examined further. It could be fruitful to investigate children who show difficulties with lexical-semantic skills. Finally, measuring statistical learning online could be beneficial for both group comparisons as well as studying individual differences.

Data, code and materials availability statement

R scripts, data and materials are available on FigShare: 10.21942/uva.c.5174660 (reserved DOI).

Ethics statement

Ethics approval was obtained from the ethics committee of the University of Amsterdam. The caretakers of all participants gave informed written consent before the participant took part in the study.

Authorship and contributor ship

Iris Broedelet: conceptualization, methodology, formal analysis, investigation, data curation, writing – original draft, visualization, project administration, funding acquisition; Paul Boersma: conceptualization, methodology, formal analysis, writing – review and editing, supervision; Judith Rispens: conceptualization, methodology, writing – review and editing, supervision.

Acknowledgements

We offer our sincere gratitude to all children who participated and their caretakers, all speech therapists and teachers from VierTaal, The Royal Auris Group and Pento for their extensive help on facilitating children's participation, as well as the staff members from the primary schools that participated in our research. We would also like to thank our test-assistant Maayke Sterk and our lab technician Dirk Jan Vet.

References

Aguilar, J. M., & Plante, E. (2014). Learning of grammar-like visual sequences by adults with and without language-learning disabilities. *Journal of Speech, Language, and Hearing Research, 57*(4), 1394–1404. https://doi.org/10.1044/2014_JSLHR-L-13-0124

Aguilar, J. M., Plante, E., & Sandoval, M. (2017). Exemplar Variability Facilitates Retention of Word Learning by Children With Specific Language Impairment. *Language, Speech, and Hearing Services in Schools*, 1–13. https://doi.org/10.1044/2017_LSHSS-17-0031

Alt, M., & Plante, E. (2006). Factors that influence lexical and semantic fast mapping of young children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 49*(5), 941–954. https://doi.org/10.1044/1092-4388(2006/068)

Altvater-Mackensen, N., Jessen, S., & Grossmann, T. (2017). Brain responses reveal that infants' face discrimination is guided by statistical learning from distributional information. *Developmental Science, 20*(2). https://doi.org/10.1111/desc.12393

Arciuli, J., & Conway, C. M. (2018). The promise—and challenge—of statistical learning for elucidating atypical language development. *Current Directions in Psychological Science, 27*(6), 492–500. https://doi.org/10.1177/0963721418779977

Arciuli, J., & Simpson, I. C. (2011). Statistical learning in typically developing children: the role of age and speed of stimulus presentation. *Developmental Science, 14*(3), 464–73. https://doi.org/10.1111/j.1467-7687.2009.00937.x

Arnon, I. (2019). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods*, 1–14. https://doi.org/10.3758/s13428-019-01205-5

Bates D., Mächler M., Bolker B., & Walker S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beckage, N., Smith, L., & Hills, T. (2010). Semantic network connectivity is related to vocabulary growth rate in children. In *Proceedings of the Cognitive Science Society* (Vol. 32, pp. 2769–2774).

Bishop, D. V. (2006). What causes specific language impairment in children? *Current Directions in Psychological Science, 15*(5), 217–221. https://doi.org/10.1111/j.1467-8721.2006.00439.x

Brackenbury, T., & Pye, C. (2005). Semantic deficits in children with language impairments. *Language, Speech, and Hearing Services in Schools, 36*, 5–16. https://doi.org/10.1044/0161-1461(2005/002)

Broedelet, I., Boersma, P., & Rispens, J. (2022). School-aged children learn novel categories on the basis of distributional information. *Frontiers in Psychology*, 12. https://doi.org/10.3389/fpsyg.2021.799241

Chládková, K., Boersma, P., & Escudero, P. (2022). Unattended distributional training can shift phoneme boundaries. *Bilingualism: Language and Cognition*, 1–14. https://doi.org/10.1017/S1366728922000086

Collisson, B. A., Grela, B., Spaulding, T., Rueckl, J. G., & Magnuson, J. S. (2015). Individual differences in the shape bias in preschool children with specific language impairment and typical language development: Theoretical and clinical implications. *Developmental Science, 18*(3), 373–388. https://doi.org/10.1111/desc.12219

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*(3), 356–371. https://doi.org/10.1016/j.cognition.2009.10.009

Dockrell, J. E., Messer, D., & George, R. (2001). Patterns of naming objects and actions in children with word finding difficulties. *Language and Cognitive Processes, 16*(2-3), 261–286. https://doi.org/10.1080/01690960042000030

Dockrell, J. E., Messer, D., George, R., & Ralli, A. (2003). Beyond naming patterns in children with WFDs—Definitions for nouns and verbs. *Journal of Neurolinguistics, 16*(2), 191–211. https://doi.org/10.1016/S0911-6044(02)00012-X

Drljan, B., & Vukovic, M. (2019). Comparison of lexical-semantic processing in children with developmental language disorder and typically developing peers. *Govor, 36*(2), 119–138. https://doi.org/10.22210/govor.2019.36.07

Eadie, P., Conway, L., Hallenstein, B., Mensah, F., McKean, C., & Reilly, S. (2018). Quality of life in children with developmental language disorder. *International Journal of Language & Communication Disorders*, 1–12. https://doi.org/10.1111/1460-6984.12385

Ebert, K. D., & Kohnert, K. (2011). Sustained attention in children with primary language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*. https://doi.org/10.1044/1092-4388(2011/10-0231)

Ellis, E. M., Gonzalez, M. R., & Deák, G. O. (2014). Visual prediction in infancy: what is the association with later vocabulary? *Language Learning and Development*, *10*(1), 36–50. https://doi.org/10.1080/15475441.2013.799988

Escudero, P., Benders, T., & Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *The Journal of the Acoustical Society of America*, *130*(4), EL206–EL212. https://doi.org/10.1121/1.3629144

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *52*(2), 321–335. https://doi.org/10.1044/1092-4388(2009/07-0189)

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*(4), 513. https://doi.org/10.1017.S0142716406060383

Gathercole, S. E., Hitch, G. J., Service, E., & Martin, A. J. (1997). Phonological short-term memory and new word learning in children. *Developmental Psychology*, *33*(6), 966–79. https://doi.org/10.1037/0012-1649.33.6.966

Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological Memory and Vocabulary Development During the Early School Years: A Longitudinal Study. *Developmental Psychology*, *28*(5), 887–898. https://doi.org/10.1037/0012-1649.28.5.887

Goujon, A. (2011). Categorical implicit learning in real-world scenes: evidence from contextual cueing. *Quarterly Journal of Experimental Psychology (2006)*, *64*(5), 920–41. https://doi.org/10.1080/17470218.2010.526231

Haebig, E., Saffran, J. R., & Ellis Weismer, S. (2017). Statistical word learning in children with autism spectrum disorder and specific language impairment. *Journal of Child Psychology and Psychiatry*, 1251–1263. https://doi.org/10.1111/jcpp.12734

Hamrick, P., Lum, J. A., & Ullman, M. T. (2018). Child first language and adult second language are both tied to general-purpose learning systems. *Proceedings of the National Academy of Sciences*, 1–6. https:///doi.org/10.1073/pnas.1713975115

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, *23*(1), 65–94. https://doi.org/10.1177/0267658307071601

Hedenius, M., Persson, J., Tremblay, A., Adi-Japha, E., Veríssimo, J., Dye, C. D., … Bruce Tomblin, J. (2011). Grammar predicts procedural learning and consolidation deficits in children with Specific Language Impairment. *Research in Developmental Disabilities*, 2362–2375. https://doi.org/10.1016/j.ridd.2011.07.026

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child development*, *74*(5), 1368-1378. https://doi.org/10.1111/1467-8624.00612

Hsu, H. J., & Bishop, D. V. (2010). Grammatical difficulties in children with specific language impairment: Is learning deficient? *Human Development*, *53*(5), 264–277. https://doi.org/10.1159/000321289

Hsu, H. J., Tomblin, J. B., & Christiansen, M. H. (2014). Impaired statistical learning of non-adjacent dependencies in adolescents with specific language impairment. *Frontiers in Psychology*, *5*, 1–10. https://doi.org/10.3389/fpsyg.2014.00175

Junge, C., van Rooijen, R., & Raijmakers, M. (2018). Distributional Information Shapes Infants' Categorization of Objects. *Infancy*, *23*(6), 917–926. https://doi.org/10.1111/infa.12258

Kachergis, G., Yu, C., & Shiffrin, R. M. (2014). Cross-situational word learning is both implicit and strategic. *Frontiers in Psychology*, *5*, 1–10. https://doi.org/10.3389/fpsyg.2014.00588

Kan, P. F., & Windsor, J. (2010). Word learning in children with primary language impairment: a meta-analysis. *Journal of Speech, Language, and Hearing Research : JSLHR*, *53*(3), 739–756. https://doi.org/10.1044/1092-4388(2009/08-0248)

Kapa, L. L., & Erikson, J. A. (2020). The relationship between word learning and executive function in preschoolers with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research*, *63*(7), 2293–2307. https://doi.org/10.1044/2020_JSLHR-19-00342

Kaufman, S. B., Deyoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*(3), 321–340. https://doi.org/10.1016/j.cognition.2010.05.011

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

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development, 87*(1), 184–193. https://doi.org/10.1111/cdev.12461

Kraemer, H. C., & Blasey, C. M. (2004). Centring in regression analyses: a strategy to prevent errors in statistical inference. *International Journal of Methods in Psychiatric Research, 13*(3), 141–51. https://doi.org/10.1002/mpr.170

Lahey, M., & Edwards, J. (1999). Naming errors of children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 42*(1), 195–205. https://doi.org/10.1044/jslhr.4201.195

Lammertink, I., Boersma, P., Rispens, J., & Wijnen, F. (2020). Visual statistical learning in children with and without DLD and its relation to literacy in children with DLD. *Reading and Writing*, 1–33. https://doi.org/10.1007/s11145-020-10018-4

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2017). Statistical Learning in Specific Language Impairment: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, 1–13. https://doi.org/10.1044/2017_JSLHR-L-16-0439

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2019). Children with Developmental Language Disorder have an auditory verbal statistical learning deficit: evidence from an online measure. *Language Learning*. https://doi.org/10.1111/lang.12373

Leonard, L. B. (2014). *Children with specific language impairment*. MIT press.
Leonard, L. B., Nippold, M. A., Kail, R., & Hale, C. A. (1983). Picture naming in language-impaired children. *Journal of Speech and Hearing Research, 26*(4), 609–15. https://doi.org/10.1044/jshr.2604.609

Lian, A. (2017). Statistical learning and developmental language impairments. *Scandinavian Psychologist, 4*, 1–16. https://doi.org/10.1057/978-1-137-58746-6_8

Lukács, A., & Kemény, F. (2014). Domain-general sequence learning deficit in specific language impairment. *Neuropsychology, 28*(3), 472. https://doi.org/10.1037/neu0000052

Lum, J. A., & Conti-Ramsden, G. (2013). Long-term memory: A review and meta-analysis of studies of declarative and procedural memory in specific language impairment. *Topics in Language Disorders, 33*(4), 282–297. https://doi.org/10.1097/01.TLD.0000437939.01237.6a

Lum, J. A., Conti-Ramsden, G., Morgan, A. T., & Ullman, M. T. (2014). Procedural learning deficits in specific language impairment (SLI): a meta-analysis of serial reaction time task performance. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 51*, 1–10. https://doi.org/10.1016/j.cortex.2013.10.011

Mainela-Arnold, E., & Evans, J. L. (2014). Do statistical segmentation abilities predict lexical-phonological and lexical-semantic abilities in children with and without SLI? *Journal of Child Language, 41*(2), 327–351. https://doi.org/10.1017/S0305000912000736

Mainela-Arnold, E., Evans, J. L., & Coady, J. A. (2010). Explaining lexical-semantic deficits in specific language impairment: The role of phonological similarity, phonological working memory, and lexical competition. *Journal of Speech, Language, and Hearing Research, 53*(6), 1742–1756. https://doi.org/10.1044/1092-4388(2010/08-0198)

Marinellie, S. A., & Johnson, C. J. (2002). Definitional skill in school-age children with specific language impairment. *Journal of Communication Disorders, 35*(3), 241–259. https://doi.org/10.1016/S0021-9924(02)00056-4

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science, 11*(1), 122–134. https://doi.org/10.1111/j.1467-7687.2007.00653.x
Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101–B111. https://doi.org/10.1016/s0010-0277(01)00157-3

McGregor, K. K. (1997). The nature of word-finding errors of preschoolers with and without word-finding deficits. *Journal of Speech, Language, and Hearing Research, 40*(6), 1232–1244. https://doi.org/10.1044/jslhr.4006.1232

McGregor, K. K., & Appel, A. (2002). On the relation between mental representation and naming in a child with specific language impairment. *Clinical Linguistics & Phonetics, 16*(1), 1–20. https://doi.org/10.1080/02699200110085034

McGregor, K. K., Berns, A. J., Owen, A. J., Michels, S. A., Duff, D., Bahnsen, A., J. & Lloyd, M. (2012). Associations between syntax and the lexicon among children with or without ASD and language impairment. *Journal of Autism and Developmental Disorders, 42*(1), 35–47. https://doi.org/10.1007/s10803-011-1210-4

McGregor, K. K., Newman, R. M., Reilly, R. M., & Capone, N. C. (2002). Semantic representation and naming in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*. https://doi.org/10.1044/1092-4388(2002/081)

McGregor, K. K., Oleson, J., Bahnsen, A., & Duff, D. (2013). Children with developmental language impairment have vocabulary deficits characterized by limited breadth and depth. *International Journal of Language \& Communication Disorders, 48*(3), 307–319. https://doi.org/10.1111/1460-6984.12008

Melby-Lervåg, M., Lervåg, A., Lyster, S.-A. H., Klem, M., Hagtvet, B., & Hulme, C. (2012). Nonword-repetition ability does not appear to be a causal influence on children's vocabulary development. *Psychological Science, 23*(10), 1092–1098. https://doi.org/10.1177/0956797612443833

Metsala, J. L. (1999). Young children's phonological awareness and nonword repetition as a function of vocabulary development. *Journal of Educational Psychology, 91*(1), 3. https://doi.org/10.1037/0022-0663.91.1.3

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010). On-line individual differences in statistical learning predict language processing. *Frontiers in Psychology, 1*, 1–9. https://doi.org/10.3389/fpsyg.2010.00031

Montgomery, J. W., Magimairaj, B. M., & Finney, M. C. (2010). Working memory and specific language impairment: An update on the relation and perspectives on assessment and treatment. *American Journal of Speech-Language Pathology*. https://doi.org/10.1044/1058-0360(2009/09-0028)

Nash, M., & Donaldson, M. L. (2005). Word learning in children with vocabulary deficits. *Journal of Speech, Language, and Hearing Research, 48*(2), 439–458. https://doi.org/10.1044/1092-4388(2005/030)

Nation, K. (2014). Lexical learning and lexical processing in children with developmental language impairments. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*, 1–10. https://doi.org/10.1098/rstb.2012.0387

Newman, R., Ratner, N. B., Jusczyk, A. M., Jusczyk, P. W., & Dow, K. A. (2006). Infants' early ability to segment the conversational speech signal predicts later language development: a retrospective analysis. *Developmental Psychology, 42*(4), 643–655. https://doi.org/10.1037/0012-1649.42.4.643

Noonan, N. B. (2018). *Exploring the Process of Statistical Language Learning*.

Nunnally, J.C., & Bernstein, I.H. (1994). The Assessment of Reliability. *Psychometric Theory, 3,* 248-292.

Obeid, R., Brooks, P. J., Powers, K. L., Gillespie-Lynch, K., & Lum, J. A. (2016). Statistical learning in specific language impairment and autism spectrum disorder: A meta-analysis. *Frontiers in Psychology, 7,* 1–18. https://doi.org/10.3389/fpsyg.2016.01245

Paciorek, A., & Williams, J. N. (2015). Semantic implicit learning. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 67–88). John Benjamins.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience, 5*(4), 291–303. https://doi.org/10.1038/nrn1364

Plante, E., Gómez, R., & Gerken, L. (2002). Sensitivity to word order cues by normal and language/learning disabled adults. *Journal of Communication Disorders, 35*(5), 453–462. https://doi.org/10.1016/s0021-9924(02)00094-1

Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition, 106*(2), 665–81. https://doi.org/10.1016/j.cognition.2007.04.003

Psychology Software Tools, Inc. (2016). E-Prime 3.0. [Computer program].
Raven, J., Raven, J.C., & Court, J.H. (2003). Manual for Raven's progressive matrices and vocabulary scales. San Antonia: Harcourt Assessment.

R Core Team (2020). A language and environment for statistical computing. Vienna, Austria. https://www.r-project.org/. [Computer program].

Rispens, J., & Baker, A. (2012). Nonword repetition: The relative contributions of phonological short-term memory and phonological representations in children with language and reading impairment. *Journal of Speech, Language, and Hearing Research.* https://doi.org/10.1044/1092-4388(2011/10-0263)

Rogers, L., Park, S. H., & Vickery, T. (2020). *Visual Statistical Learning Is Modulated by Arbitrary and Natural Categories.* PsyArXiv. https://doi.org/10.31234/osf.io/9ca28

Saffran, J. R. (2018). Statistical learning as a window into developmental disabilities. *Journal of Neurodevelopmental Disorders, 10*(1), 35. https://doi.org/10.1186/s11689-018-9252-y

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Sandgren, O., Salameh, E.-K., Nettelbladt, U., Dahlgren-Sandberg, A., & Andersson, K. (2020). Using a word association task to investigate semantic depth in swedish-speaking children with developmental language disorder. *Logopedics Phoniatrics Vocology*, 1–7. https://doi.org/10.1080/14015439.2020.1785001

Sanjeevan, T., & Mainela-Arnold, E. (2019). Characterizing the Motor Skills in Children with Specific Language Impairment. *Folia Phoniatrica et Logopaedica, 71*(1), 42–55. https://doi.org/10.1159/000493262

Schlichting, L. (2005). Peabody Picture Vocabulary Test-III-NL [Measurement instrument]. Amsterdam, The Netherlands: Harcourt.

Semel, E., Wiig, E., & Secord, W. (2010). Clinical evaluation of language fundamentals: Dutch Version (W. Kort, E. Compaan, M. Schittekatte, & P. Dekker, Trans.; Third Edition.) [Measurement instrument]. Amsterdam, The Netherlands: Pearson.

Shafto, C. L., Conway, C. M., Field, S. L., & Houston, D. M. (2012). Visual sequence learning in infancy: Domain-general and domain-specific associations with language. *Infancy, 17*(3), 247–271. https://doi.org/10.1111/j.1532-7078.2011.00085.x

Sheng, L., & McGregor, K. K. (2010). Lexical-semantic organization in children with specific language impairment. *Journal of Speech, Language, and Hearing Research : JSLHR, 53*(1), 146–159. https://doi.org/10.1044/1092-4388(2009/08-0160)

Siegelman, N. (2020). Statistical learning abilities and their relation to language. *Language and Linguistics Compass, 14*(3), e12365. https://doi.org/10.1111/lnc3.12365

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B, 372*(1711), 1–10. http://doi.org/10.1098/rstb.2016.0059

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods, 49*(2), 418–432. https://doi.org/10.3758/s13428-016-0719-z

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2017). Redefining "learning" in statistical learning: what does an online measure reveal about the assimilation of visual regularities? *Cognitive Science*, 1–36. https://doi.org/10.1111/cogs.12556

Singh, L., Steven Reznick, J., & Xuehua, L. (2012). Infant word segmentation and childhood vocabulary development: a longitudinal analysis. *Developmental Science*, *15*(4), 482–95. https://doi.org/10.1111/j.1467-7687.2012.01141.x

Snowling, M.J., Hayiou-Thomas, M.E., Nash, H.M., & Hulme, C. (2020). Dyslexia and Developmental Language Disorder: comorbid disorders with distinct effects on reading comprehension. *Journal of Child Psychology and Psychiatry, 61*(6), 672-680. https://doi.org/10.1111/jcpp.13140.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3), 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010

Sociaal en Cultureel Planbureau (2018). *Statusscores 2016* [report Social and Cultural Planning Bureau]. No longer available.

Spencer, M., Kaschak, M. P., Jones, J. L., & Lonigan, C. J. (2015). Statistical learning is related to early literacy-related skills. *Reading and Writing*, *28*(4), 467–490. https://doi.org/10.1007/s11145-014-9533-0

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology*, *126*, 395–411. http://doi.org/10.1016/j.jecp.2014.06.0030022-0965/?

Tomblin, J. B., Mainela-Arnold, E., & Zhang, X. (2007). Procedural learning in adolescents with and without specific language impairment. *Language Learning and Development*, *3*(4), 269–293. https://doi.org/10.1080/15475440701377477

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology. General*, *134*(4), 552–64. https://doi.org/10.1037/0096-3445.134.4.552

Ullman, M. T. (2016). The declarative/procedural model: a neurobiological model of language learning, knowledge and use. In G. Hickock & S. A. Small (Eds.), *The neurobiology of language* (pp. 953–968). Elsevier.

Ullman, M. T., & Pierpont, E. I. (2005). Specific language impairment is not specific to language: The procedural deficit hypothesis. *Cortex, 41*(3), 399–433. https://doi.org/10.1016/S0010-9452(08)70276-4

Vandermosten, M., Wouters, J., Ghesquière, P., & Golestani, N. (2019). Statistical learning of speech sounds in dyslexic and typical reading children. *Scientific Studies of Reading, 23*(1), 116–127. https://doi.org/10.1080/10888438.2018.1473404

Wanrooij, K., Boersma, P., & Benders, T. (2015). Observed effects of "distributional learning" may not relate to the number of peaks. A test of "dispersion" as a confounding factor. *Frontiers in Psychology*, *6*, 1341. https://doi.org/10.3389/fpsyg.2015.01341

Wanrooij, K., Boersma, P., & van Zuijen, T. L. (2014). Fast phonetic learning occurs already in 2-to-3-month old infants: an ERP study. *Frontiers in Psychology*, *5*, 77. https://doi.org/10.3389/fpsyg.2014.00077

Westby, C. (2019). Depression in Children With Developmental Language Disorders. *Word of Mouth*, *30*(4), 1–4. https://doi.org/10.1177/1048395019833703

Wu, R., Gopnik, A., Richardson, D. C., & Kirkham, N. Z. (2011). Infants learn about objects from statistics and people. *Developmental Psychology*, *47*(5), 1220–9. https://doi.org/10.1037/a0024023

Wu, R., Gopnik, A., Richardson, D., & Kirkham, N. (2010). Social cues support learning about objects from statistics in infancy. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32).

Younger, B. A. (1985). The segregation of items into categories by ten-month-old infants. *Child Development*, *56*(6), 1574–83.

Yu, C., & Smith, L. B. (2011). What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science*, *14*(2), 165–180. https://doi.org/10.1111/j.1467-7687.2010.00958.x

License

Pauses matter: Rule-learning in Children

Anika van der Klis*
Utrecht University, the Netherlands

Rianne van Lieburg*
University of Antwerp, Belgium

Lisa Lai-Shen Cheng
Clara Cecilia Levelt
Leiden University, the Netherlands

* these authors contributed equally to this work

Abstract: Language learners have to both segment words and discover grammatical rules connecting those words in sentences. In adult listeners, the presence of a prosodic cue in the speech stream, for example, a pause, appears to facilitate rule-learning of non-adjacent dependencies of the form A_iXC_i (Peña et al., 2002). Only when listening to the artificial language containing pauses, could participants identify rule-words of the form $A_iA_jC_i$ or $A_iC_jC_i$, where intervening syllables were moved from A- or C-positions. Frost and Monaghan (2016) found in a similar study that participants who were tested with novel, rather than moved, intervening syllables in A_iXC_i items showed rule-learning even when the familiarisation stream contained no pauses. The present study re-examines the facilitative effect of pauses in discovering structural rules in speech in a novel population: children aged 7-11. We used the same artificial speech stimuli as Peña et al. (2002) and tested children in both a moved-syllable and novel-syllable forced-choice task. The results of 140 children show that pauses provide a facilitative effect on rule-learning – also for young learners. Regardless of syllable types, only children who listened to the familiarisation stream containing pauses chose words following the rule above chance-level.

Keywords: artificial grammar learning; statistical learning; non-adjacent dependencies; prosody; school-age children

Corresponding author(s): Rianne van Lieburg, Centre for Computational Linguistics and Psycholinguistics, University of Antwerp, Grote Kauwenberg 18, 2000 Antwerpen, Belgium. Email: rianne.vanlieburg@uantwerpen.be

ORCID ID(s): https://orcid.org/0000-0001-5024-8096

Citation: van der Klis, A., van Lieburg, R., Cheng, L.L., & Levelt, C.C. (2022). Pauses matter: Rule-learning in children. *Language Development Research, 3*(1), 44–64. http://doi.org/10.34842/2023.0466

Introduction

Language learners need to both segment words and discover grammatical rules connecting those words in sentences. Saffran et al. (1996a) demonstrated that 8-month-olds were able to segment words from running speech after a short exposure based on statistical relationships between neighbouring speech syllables. They could infer word boundaries between two syllables with a low transitional probability in the sequence (i.e., a transitional probability of 0.33 between words versus a probability of 1.0 within words). Adults have also been shown to use dips in transitional probability to infer word boundaries and to successfully extract words from a continuous speech stream (e.g., Saffran et al., 1996b; Peña et al., 2002).

Peña et al. (2002) suggested that, while statistical relationships are sufficient for speech segmentation, additional cues are needed for the detection of grammatical rules. The artificial language to which their adult participants were exposed consisted of trisyllabic sequences that followed a non-adjacent dependency (NAD) rule of the form A_iXC_i, where A_i always predicted C_i (e.g., the English progressive *is X-ing*). Hence, the transitional probability between A_i and C_i was 1.0. The within-word transitional probability (between A_i and the adjacent X or between X and the adjacent C_i) was 0.33 while between words (between the last syllable of any item and the first syllable of the next item) it was 0.5. Peña et al. (2002) showed that, when presented with a continuous speech stream, listeners only deemed test items that had appeared in the same form in the stream correct (i.e., they could segment the trisyllabic items) but deemed test items constructed by moving an A or C syllable to the X-position, resulting in $A_iA_jC_i$ or $A_iC_jC_i$, incorrect (i.e., they could not find words following the rule when there was an intervening element originating from another position in the stream). When gaps of 25-ms, which Peña et al. (2002) called "subliminal pauses", were placed between each A_iXC_i triplet (e.g., puRAki-pause-puLIki) in the familiarisation stream, i.e. segmenting the stream into smaller constituents, adults *were* able to extract possible words that followed the NAD rule (hereafter rule-words) containing moved syllables. This showed that participants could identify the NAD rule when – and only when – pauses were added to the otherwise identical familiarisation stream.

Peña et al. (2002) assumed that streams that are segmented by pauses relieve listeners of the task of computing probabilities to segment words, thereby giving them the chance to discover underlying rules. This hypothesis is in line with studies suggesting that prosodic cues that mark the boundaries of constituents may have a scaffolding function during language acquisition (e.g., Soderstrom et al., 2003; Morgan, 1986). Several studies showed that adults are not able to extract and generalise NADs from continuous speech streams without perceptual cues marking phrases. Similar to Peña et al. (2002), Endress and Bonatti (2007) showed that adults only preferred class-words

(of the type A_iXC_j) to part-words (of the type XCA or CAX) when listening to a familiarisation stream containing 25-ms pauses. Without these cues, participants did not show a preference for either class-words or part-words. Marchetto and Bonatti (2013, 2015) also examined children (12-month-olds and 18-month-olds) using head-turn procedures and found evidence that they could learn NADs of the type A_iXC_i, but only when listening to a stream containing either 25-ms or 200-ms pauses. The authors proposed that the same learning mechanism used by adults might be readily available for infants - triggered by the same acoustic properties in the stream. Grama et al. (2016) examined whether other types of perceptual cues affect the learning of artificial NADs in adults. They found that performance in a forced-choice task increased when the dependent elements (i.e., A_i and C_i in A_iXC_i strings) were either acoustically enhanced or reduced in the familiarisation stream, but only when the A_iXC_i strings were also separated by pauses. This led to the hypothesis that NAD learning is easiest when the dependent elements are both perceptually distinctive *and* integrated into a prosodically natural contour (Grama et al., 2016). These behavioural results show that NAD learning in both infants and adults is enhanced when prosodic cues are present that break up the continuous speech stream into smaller constituents containing the rule (i.e., A_iXC_i strings). These smaller constituents may play a facilitative role to learners extract rules.

The processing of NADs in the brain has been studied too, using electroencephalography (EEG). Mueller et al. (2008) found that adult participants showed different event-related potential (ERP) patterns when listening to a stream containing pauses, in addition to an increase in correct responses by 30% in a condition with pauses compared to a condition without pauses. In the condition with pauses, participants showed an additional positivity in their responses, which the authors interpreted as reflecting more controlled, attention-guided mechanisms. De Diego-Balaguer et al. (2015) also examined ERPs in adults while they listened to different artificial speech streams containing trisyllabic items with and without 25-ms pauses in between them. Their results showed that pauses altered electrophysiological responses to the stream. In the stream without pauses, the amplitude N1 component increased at syllable onsets, which indicates that participants pay attention to them for the sake of locating word boundaries. Pauses reduced the mean amplitude of the N1 component in the first syllable of the trisyllabic items, which may indicate that participants segment the stream by means of the pauses, and no longer need to orient to the syllable onset for the location of the word boundaries. Behavioural results of this study also showed that while participants were indeed better at segmenting words when the continuous speech stream contained pauses, these pauses did not improve rule learning (de Diego-Balaguer et al., 2015). The findings of these studies corroborate Peña et al.'s (2002) hypothesis that the availability of perceptual cues relieve listeners of the speech segmentation task and alter processing of the speech stream, but not their necessity for rule-learning.

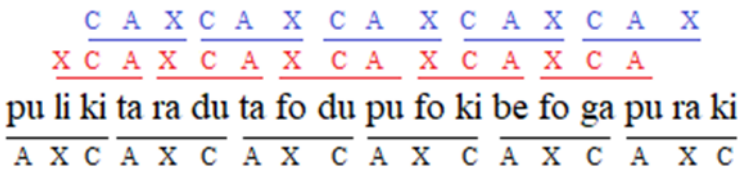


Figure 1. ***A short excerpt of the familiarisation stream used by Peña et al. (2002) containing six AXC rule-words (black) and ten part-words of the type XCA (red) and CAX (blue).***

Nonetheless, the interpretation that statistical relationships alone do not suffice for rule-learning is heavily debated in the literature (see Perruchet et al., 2004; Endress & Mehler, 2009; Frost & Monaghan, 2016). Peña et al. (2002) concluded that participants did not learn the rule in the non-pause condition because participants significantly chose part-words (i.e., of the form XCA or CAX) over rule-words (i.e., of the form $A_iA_jC_i$ or $A_iC_jC_i$). However, Frost and Monaghan (2016) pointed out that, even though infrequently, part-words had appeared in the familiarisation stream exactly as such, while rule-words containing moved syllables, such as *pubeki*, had not. The artificial language stream consists of many adjacent words, and part-words were formed by syllables that span word boundaries, as illustrated in Figure 1. This excerpt of only six rule-words in fact contains ten part-words. Participants were thus forced in the test to choose between part-words, that had appeared in the familiarisation stream, and rule-words with moved syllables, that had never occurred as such. Preferring rule-words over part-words, then, requires not only identification of the structural generalisation, but also suppression of learned (or encountered) syllable sequences. Frost and Monaghan (2016) used the same artificial language used by Peña et al. (2002) in their study and created test items where the intervening elements were either moved (from A or C positions) or completely novel. Their 10.5-min long familiarisation stream did not contain pauses or any other prosodic cues. Adults selected rule-words rather than part-words significantly above chance, but only when the test items contained novel, rather than moved, intervening elements (M = .693, p < .001). They therefore concluded, in line with Endress and Mehler (2009) and Perruchet et al. (2004), that the pause used by Peña et al. (2002) only served as an additional cue, increasing the saliency of the positions of individual syllables, rather than relieving listeners from the segmentation task. The result from Frost and Monaghan (2016) further questions the actual role of prosodic cues that mark constituent boundaries in rule-learning. The experiments reported in the present paper aim to inform this debate.

Currently, it is not known whether school-aged children show performance similar to adults, because this is an underrepresented age group in the rule-learning

literature. Research has yet to find out whether school-aged children, like adults, can generalise NAD-rules over test items with novel syllables. In Soderstrom et al. (2007), 16-month-old infants could only generalise NADs in a natural language to novel nonsense stems (e.g., vod teebs) if they were first presented with familiar stems (e.g., dog runs). The authors hypothesised that infants had been distracted by the presence of unfamiliar words in the stimuli. Similarly, Grama and Wijnen (2018) showed, using an artificial language paradigm, that while 18-month-olds do have abstract knowledge of A_iXC_i strings after exposure, they cannot generalise the NADs when there are novel intervening syllables. Novel items may actually draw children's attention away from the dependency, yielding hindering, rather than facilitating effects on rule learning. These results are in contrast with the findings by Frost and Monaghan (2016) for adults, for whom the use of novel X-syllables did not hinder the ability to generalise the NADs. The present study is the first to assess artificial rule-learning using novel stimuli in school-age children.

It is also not known if school-age children can successfully learn NAD-rules during passive listening. Mueller et al. (2012) found that adults were less successful than infants in NAD learning under passive conditions when measuring ERPs. In an ERP experiment with 2- and 4-year-olds, Mueller et al. (2019) showed that passive learning of NADs, in an artificial language with pauses, declined between 2 and 4 years of age. This is linked to maturation of the Prefrontal Cortex (PFC), which is completed around the age of 7 and involves a switch to a different, more adult-like learning mechanism (Skeide & Friederici, 2016). Similarly, van der Kant et al. (2020), using functional near-infrared spectroscopy (fNIRS), found evidence that only 2-year-olds, but not 3-year-olds could detect linguistic NAD violations during passive listening. In a recent study using ERPs, Paul et al. (2021) examined children between 1 and 3 years old, and although all children showed evidence of learning NADs in a foreign language, there was a gradual decrease in the strength of this evidence across these ages. It may therefore be possible that during development, there is a decline in the ability to learn through passive listening when there are no additional cues that mark the NADs. This suggests that children aged 7 to 11 may not be as successful as infants in detecting NADs during passive listening.

Previous studies have shown that adults outperform children when NAD learning is assessed using a task that requires more declarative knowledge, such as a grammatical judgement task, even when they do not receive instructions prior to listening to the speech stream. Ferman and Karni (2010) examined artificial grammar learning in adults, 12-year-olds, and 8-year-olds. Adults outperformed both groups of children, and 12-year-olds outperformed 8-year-olds. This was reflected in higher accuracy as well as shorter response times in both a grammatical judgement and a production task. Ojima and Okanoya (2020) tested adults and children aged 5 to 12 on centre-embedding learning in an artificial A^nB^n grammar which also generates NADs.

They found that the majority of the adults could generalise the rules to novel stimuli in a go/no-go task, indicating that they had learned the grammar. However, only about a quarter of the children in their study succeeded in this. The authors suggested that failure in this task by all the other children is due to memory constraints, not due to rule-learning deficits. Lammertink et al. (2019) examined NAD learning in children aged 5 to 8 (M = 7.3) years old. They did find evidence for sensitivity to NADs in online reaction times, but above-chance performance was not found in an offline forced-choice task. The children in their study did not receive explicit instructions, but their task did require a certain level of attention to the stimuli. The authors argued that this grammatical judgement task required more metalinguistic awareness and attention, which is more difficult for children compared to adults. Marimom et al. (2021) ran a similar experiment with children aged 3 to 8 (M = 6.2) years old, but they used a stem completion task instead of a forced-choice task. They found evidence of learning in reaction times, and above-chance performance at the group level during the stem completion task. The results furthermore showed faster reaction times for older children, although their accuracy scores did not increase. Importantly, Marimom et al. (2021) added 20-ms pauses at the beginning of each AXB stimulus as well as a longer pause between the A-element and X-element, which may have enhanced children's performance in this study. Schaadt et al. (2020), in a study with 7-year-olds, found no significant above-chance performance at the group level on a grammatical judgement task (choosing between *correct* or *incorrect*), after familiarisation with short sentences - separated by pauses - containing NADs in a natural foreign language. However, accompanying ERP data did show, both at the group and at the individual level, that especially after a retention period of one night's sleep, a representation of the NAD had been built. Children can implicitly recall NADs, as shown by more negative ERP responses to NAD violations, but they do not show explicit knowledge in the grammatical judgement task. The authors concluded that their grammatical judgement task was still too difficult for children of this age, and that they might have been able to show an effect of learning in a forced-choice task. The present study uses this more suitable task and aims to add to our understanding of NAD learning in this age group.

In our study, we investigated the performance of 7- to 11-year-old children who were tested on both moved and novel intervening syllables in the AXB test items. We expected children above the age of 7 to have switched to a more adult-like mechanism (Skeide & Friederici, 2016) that benefits both from additional segmentation cues in the speech stream (e.g., Peña et al., 2002; Grama & Wijnen, 2018) and from a task which requires more metalinguistic knowledge to guide their attention to the NADs (e.g., Pacton & Perruchet, 2008; Bialystok, 1986) compared to younger children. In the first experiment, we used the same A_iXC_i language as in Peña et al. (2002) and created test items using moved syllables of the form $A_iA_jC_i$ or $A_iC_jC_i$ as intervening syllables. In line with Peña et al. (2002), we expect better learning of the NADs when pauses

were present in the familiarisation stream. Pauses segment the stream into constituents, which draws more attention to the dependent elements on constituent boundaries. This could help children discover rules. However, this experiment does not specifically address the question of whether a segmented stream also facilitates the discovery of NAD-rules when participants are presented with novel intervening syllables, or whether using novel elements alone is sufficient to draw children's attention to underlying rules. In our second experiment, we tested a new group of children and used novel syllables as intervening syllables in the test items. Here there are two possible outcomes; either the novel intervening syllables are sufficient for drawing children's attention to the underlying rule without needing other segmentation cues, as Frost and Monaghan (2016) found for adults, or the novel intervening syllables end up hampering children's ability to generalise, as was found for infants (e.g., Soderstrom et al. 2007; Grama and Wijnen, 2018). If the presence of pauses results in more successful learning in the novel-syllable task, this constitutes more precise evidence for the facilitative effect of pauses. By pitting moved syllables against novel syllables, we can get a deeper understanding of the effect of pauses in artificial rule-learning in school-age children.

Method

Participants

For the first experiment, we aimed to collect as many data as we could within the duration of one semester[1]. We tested 92 children (55 boys, 37 girls[2]) between 7- and 11-years-old (M = 8.55, SD = 1.18). For the second experiment, we tested a new group, again within the duration of one semester, collecting data from 51 children (27 boys, 24 girls) aged between 7- and 11-years-old (M = 9.04, SD = 1.06). We excluded three participants because they did not follow the test instructions properly.[3] All children were native speakers of Dutch and did not report any hearing or language-related problems. They were recruited from different primary schools in the Netherlands (Leiden and Rotterdam area) and the Leiden University Babylab participant database. The parents gave their written consent and filled out a short questionnaire providing information concerning the inclusion criteria. After participating, all children

[1] The study was funded by and conducted within the Research Traineeship Programme at Leiden University, which had a fixed duration of one year.

[2] In Experiment 1, there is an imbalance in gender ratio: we tested more boys than girls. Exploratory analyses did not reveal any effect of gender in either Experiment 1 or Experiment 2.

[3] We excluded two participants from Experiment 1: one in the condition with pause (18 correct responses) and one in the condition without pauses (19 correct responses). One participant in the without-pauses condition was excluded from Experiment 2 (14 correct responses).

received a monetary compensation of five euros.

Previous studies reported strong effect sizes with adults: Frost and Monaghan (2016) report a Cohen's *d* of 1.2 on the novel syllable task testing 18 participants. Because children usually show much more variability in performance, we used two-thirds of this factor size for our power estimate: $d = (2/3 * 1.2) = 0.8$, which corresponds to an odds ratio of 4.3.[4] We used the WebPower package (Zhang & Yuan, 2018) to determine the minimum sample size required to have at least 80% power. Results indicated that the minimum sample size should be $n \geq 43$. In both our experiments our sample sizes exceeded this number.

Materials

The familiarisation stream consisted of a ten-minute long sequence of syllables created with the "Female 5" French voice[5] of the speech synthesiser Praat (Boersma & Weenink, 2018). We used the same A_iXC_i language and the same syllables as in Peña et al. (2002). The A_iXC_i dependencies were *pu*X*ki*, *ta*X*du* and *be*X*ga*. The X-syllables were *li*, *ra* and *fo,* leading to 9 different trisyllabic items: *puliki, puraki, pufoki, talidu, taradu, tafodu, beliga, beraga,* and *befoga.* It should be noted that the A_i and C_i syllables involved plosives, and the X syllables continuants, which resembles natural language (cf. Frost and Monaghan, 2016). We pseudorandomized the order of the different A_iXC_i sequences ("words") according to the same criteria as in Peña et al. (2002). Each trisyllabic item occurred a hundred times in the stream. Two subsequent items never started with the same syllable. The X-syllable always differed between two subsequent items. The transitional probability between A_i and C_i was 1.0, the within-word transitional probability (between A_i and the adjacent X or X and the adjacent C_i) was 0.33 and the between-words transitional probability (between the last syllable of any item and the first syllable of the next item) was 0.5. We created two versions of the familiarisation stream: one containing a 10-ms[6] pause between each trisyllabic item and one without such pauses, leaving them completely identical otherwise. We

[4] Odds ratio = $e^{d \times \frac{\pi}{\sqrt{3}}}$ (see Sánchez-Meca et al. 2003).

[5] Peña et al. also used a French voice, but note that their participants were native speakers of French. We also used a French synthesiser rather than a Dutch synthesiser, because the phoneme /g/ in the used syllable /ga/ is not available in a Dutch synthesiser as Dutch only has /g/ in loanwords (meaning that the children were still familiar with /g/). Crucially, we told the children that they were going to listen to a foreign language.

[6] Peña et al. (2002) reported the use of 25-ms pauses in the familiarisation stream. Since different speech synthesisers may treat such settings differently, we measured the actual pause duration between trisyllabic items from Peña et al. using Praat and mimicked those pause durations with the speech synthesiser we used. With our speech synthesiser, generating a 10-ms pause resulted in a familiarisation stream with pauses that were comparable to the ones used by Peña et al.

used 5-second fade-in and fade-out effects, following Peña et al. (2002) and Frost and Monaghan (2016), to ensure that there was no audible first or last syllable.

In the first experiment, the forced-choice task included 36 pairs of rule-words following the A_i _C_i rule with moved intervening syllables originating from A or C positions filling the _ slot (e.g., *pubeki of the type AAC*) and part-words of the types CAX (e.g., *gapufo*) or XCA (e.g., *fogapu*). Audio files of the rule-words and part-words were created in Praat using the same synthetic voice as the familiarisation stream. In the second experiment, the forced-choice task was adapted by creating test items with novel intervening syllables (i.e., *ve, no* and *si*) that had never occurred in the familiarisation stream. Both the rule-word and the part-word contained these novel syllables instead of moved syllables. The novel syllables contained continuants, like in the X-syllables in the familiarisation stream. We used different novel syllables from Frost and Monaghan (2016) (who used *ve, zo* and *thi*), in order to only use phonemes with which the children are familiar from Dutch. The forced-choice task was both programmed and run with Praat. The script containing the forced-choice task is provided in the supplementary materials.

Procedure

The experimental procedure in both experiments consisted of a familiarisation phase followed by a test phase. First, the Dutch children listened to a short excerpt from a British English television show, and they were asked whether they recognised the language, to which the majority responded that they recognised it as being English - or at least as a foreign language they heard before. We used this to explain that they were going to listen to another language, but one that they had never heard before. We instructed them that we were going to see whether they could also recognise this new language afterwards. We did not explicitly explain that they should look for rules. The children watched a video of *Pingu* (a children's animation show) while they were presented with the familiarisation stream through over-ear headphones. Participants randomly received the familiarisation stream either with or without 10-ms pauses between the A_iXC_i items.

After the familiarisation period, the children were presented with the forced-choice task. The test started with three practice items that were not included in the analysis. All children received the same pairs of test items in random order. Participants were asked to choose the item which most likely belonged to the familiarisation language (instructed as "which one belongs to the language you just heard?"). After listening to a pair consisting of a rule-word and a part-word, two big buttons with "1" and "2" appeared on the screen. The children were instructed to select either "1" or "2" using the computer mouse to select the first or second word of the pair. They could listen to the pair one more time by clicking on a replay button. The majority of

participants immediately started selecting words from the word pairs, and we did not provide them with any additional instructions. When children were reluctant to answer during the practice phase, we reassured them that they could go with their first intuition, and that they did not have to be certain about their answer. The testing phase took about ten minutes for each child to complete. This experiment was run using Praat on Windows computers.

Coding and analysis

The responses to all test pairs were automatically coded as "correct" (i.e., the participant selected the rule-word) or "incorrect" (i.e., the participant selected the part-word) by the Praat script. This resulted in a list of 36 answers, correct or incorrect, for each participant, which we used as the binary outcome variable in a generalised linear mixed-effects model with the presence of pauses as a fixed effect predictor. Each participant also received a final score between 0 and 36 correct answers. We used this to examine whether participants scored at an above chance-level (i.e., showed a learning effect). In addition, we calculated which scores were outliers. An outlier was defined as being three times the *SD* above or below the mean. Outliers were excluded from the analyses.

We analysed the results of the two experiments separately to facilitate the comparison of the results of the first experiment to the results of Peña et al. (2002), and those of the second experiment to the results of Frost and Monaghan (2016). In addition, we performed a joint analysis to further assess the relationship between the use of moved or novel syllables and pauses in NAD learning.

Results

Separate analyses

In the first experiment using test items constructed with moved intervening syllables, we analysed the results of 91 children. Children who listened to the stream without pauses (n = 46) had an average score of 16.31 correct responses (45.6%, *SD*: 3.96), whereas those who received the stream with pauses (n = 45) had an average score of 19.20 (53.3%, *SD*: 4.47) correct responses.

The responses were compared between the two groups by fitting a binomial generalised linear mixed model (R-package lme4, Bates et al., 2012). The presence of pauses in the familiarisation stream and children's age in months were included as fixed effects. Gender or an interaction between Pause and Age did not improve model fit so we report the model without them. We centred and scaled Age to increase convergeability. A random intercept was included for participants which significantly

improved model fit. The *p*-values were obtained by using the package lmerTest (Kuznetsova et al., 2017). We used the jtools package (Long, 2020) to calculate exponentiated coefficients (i.e., odds ratios).

The significant negative intercept indicates that the responses of participants who were exposed to the familiarisation stream without pauses were more often incorrect than correct. Significantly more items were answered correctly if the participant had received the familiarisation stream with pauses ($p < .01$). In addition, scores improved upon increasing age ($p < .05$). When test items contained moved intervening syllables, children who received the stream with pauses were 1.32 times more likely to give a correct answer than children who received the stream without pauses. For both conditions, we also compared the proportion of correct responses to chance-level (50%, which equals a mean of 18 correct responses) in a one-tailed *z*-test[7]. The group that received the familiarisation stream without pauses did not perform above chance-level ($p = .99$), whereas the group that listened to the familiarisation stream with pauses did ($p < .01$, $d = 0.268$).

Table 1. ***Results of the generalised linear mixed model of the first experiment using moved X-syllables (n = 3276, log-likelihood = -2245.0)***

Predictor	Estimate	Exponent. Estimate	*SE*	Wald *Z*	*p*-value
(Intercept)	-0.14	0.87	0.07	-2.00	0.05
Pause	0.27	1.32	0.10	2.74	0.01
Age	0.11	1.11	0.05	2.18	0.03

In the second experiment, using test items constructed with novel intervening syllables, the results of 49 participants were analysed. Children who listened to the familiarisation stream without pauses ($n = 24$) had an average of 18.25 (50.7%, *SD*: 3.88) correct responses, whereas children who received the familiarisation stream with pauses ($n = 25$) had an average of 19.62 (54.5%, *SD*: 4.59) correct responses. One participant, who was exposed to the familarisation stream with pauses, had an outlier score of 33 correct responses. After excluding this participant[8], the group of children

[7] We chose to perform a one-tailed z-test, because we tested for a positive learning effect. We report on the two-tailed z-tests of all our statistical comparisons against chance level in the Supplementary materials. The levels of significance remain the same, except that the group in the moved-syllables experiment that received the familiarisation stream without pauses scored significantly below chance-level ($p < .001$).

[8] The statistical analyses including the outlier are reported in the Supplementary materials. Our conclusions are not altered by this inclusion.

who received the familiarisation stream had an average of 19.20 (53.6%, *SD*: 3.76) correct responses.

We compared the responses of the two groups by fitting a binomial generalised linear mixed model (R-package lme4, Bates et al., 2012). Neither the presence of Pause (= 0.95, p = .33) nor Age (centred and scaled) (= 0.78, p = .38) nor their interaction (χ = 0.57, p = .45) improved model fit when using novel intervening syllables in the test items. The model was only improved by adding a random intercept for participants to the null model. The fixed factors do not contribute beyond the random effect to explain differences in the number of correct responses. Again, we also compared the proportion of correct responses to chance-level per condition in a one-tailed z-test. The group that was exposed to the familiarisation stream without pauses did not perform above chance-level (p = .35), whereas the group that was presented with the familiarisation stream with pauses did (p = .001, d = 0.319).

Joint analysis

To examine a possible interaction between the moved- and novel- syllable conditions, we further analysed the results by performing a joint analysis of both experiments (n = 139). We built up the model by adding a random intercept for participants which significantly improved model fit. Then, we added fixed effects step by step. We found significant improvements of fit when adding Pause and Age. An interaction between Pause and Age did not significantly improve model fit. Neither Gender nor Syllable Type improved fit. We report the final model with fixed effects of Pause and Age and a random intercept for participants in Table 2.

The intercept represents the log-odds for a correct response in the condition of exposure to the familiarisation stream without pauses and the forced-choice task of test items with moved intervening syllables. Significantly more items were answered correctly if the participant had received the familiarisation stream with pauses (p < .01). In addition, there was a positive effect of age (p < .01). Across both experiments, children who listened to a stream with pauses were 1.25 times more likely to give a correct answer compared to children who listened to the stream without pauses, regardless of syllable type.

Table 2. ***Results of the generalised linear mixed model of the joint analysis (n = 5004, log-likelihood = -3438.7)***

Predictor	Estimate	Exponent. Estimate	*SE*	Wald *Z*	*p*-value
(Intercept)	-0.10	0.91	0.05	-1.76	0.08
Pause	0.22	1.25	0.08	2.88	0.007
Age	0.11	1.11	0.04	2.71	0.004

In our model, we compared the results of the different groups but not the learning effect per se, i.e., scoring higher than chance-level. In the separate analyses of both groups, we found above chance performance in the condition with pauses, but not in the condition without pauses. Overall, participants did not score above chance (50%) (M = 18.13, SD = 4.23, n = 139, p = .31 in a one-tailed z-test). However, like in the analyses of the individual experiments, a one-tailed z-test showed that the group that listened to the familiarisation stream without pauses did not perform above chance (M = 17.04, SD = 4.02, n = 70, p > .99), whereas the group exposed to the familiarisation stream with pauses did (M = 19.23, SD = 4.18, n = 69, p = <.001, d = 0.294).

The one-tailed z-test uses mean scores. However, we also looked at individual scores (see Figure 2). For an individual score significantly higher than chance-level, at least 24 out of 36 items should be correct. Note that 'significantly higher than chance-level' means a score higher than 95% of the scores in a binomial distribution with a probability of success of 50%. The probability of a score of 24+ correct responses is 3.26%[9]. Under a binomial distribution, we would therefore expect 5 out of 140 participants to have 24+ correct items. However, 15 participants (11.4%) turned out to have 24+ correct responses, 12 of which received the familiarisation stream with pauses. Of the 3 participants who were exposed to the familiarisation stream without pauses, 2 were in the moved intervening syllable condition (4.3%) and 1 was in the novel intervening syllable condition (4.2%). In the conditions with pauses, many more participants than expected scored above chance-level (X^2 (1, N = 69) = 8.49, p < .01)). In the conditions without pauses, the number of participants that scored above chance-level was not higher than expected (X^2 (1, N = 140) = 1, p = 1)).

The experiments in this study investigated NAD learning abilities in children aged 7-11 in an artificial A_iXC_i grammar. Our study aimed to assess whether perceptual cues in the speech stream facilitate the learning of NADs in children. We examined this by pitting moved syllables against novel syllables as intervening elements in A_iXC_i strings to get a deeper understanding of the effect of pauses on NAD detection. We found that the presence of pauses indeed facilitated the detection of NADs in these young learners, regardless of syllable type in the test phase, though pauses did not guarantee that all children discovered the rule. Children who listened to the familiarisation stream with pauses chose rule-words significantly more often than part-words in a forced-choice task. When testing children using moved intervening syllables, we found a large improvement when pauses were added to the stream. The

[9] We calculated the probability of individual scores using the probability density function of the binomial distribution: $P(p,n,r) = p^r \times (1-p)^{n-r} \times \frac{n!}{r! \times (n-r)!}$, where P is the probability of a particular outcome, p is the probability of success of each trial, n is the number of trials, and r is the number of successes. We calculated at which score the cumulative probability was less than 0.05.

percentage of chosen rule-words in the forced-choice task showed a significant increase after familiarisation with the speech stream including pauses, replicating the findings by Peña et al. (2002) in a novel population: school-age children.

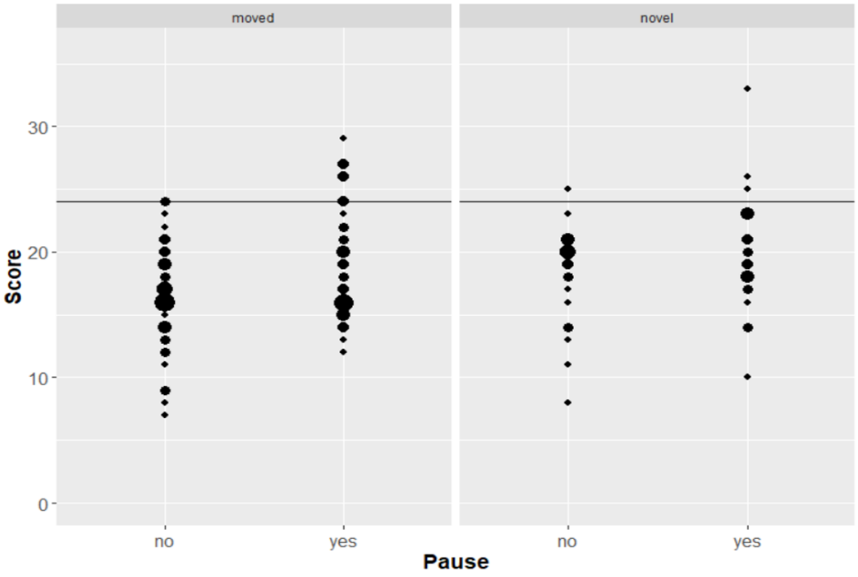


Figure 2. ***Bubble chart of number of the correct responses per condition, reference line at 24 correct responses.***

General Discussion

With regard to the effect of pauses when testing novel intervening syllables compared to the effect in the experiment with moved intervening syllables, the results are not as straightforward. There is a discrepancy between the outcomes of the separate analysis and the joint analysis. In the separate analysis of Experiment 2, there is no significant effect of Pause on scores. On the other hand, the one-tailed *z*-test showed that participants only performed significantly above chance-level (i.e., a score of 50%) when the familiarisation stream contained pauses. In addition, the joint analysis revealed no interaction between Syllable Type and Pause and a significant effect of Pause. Pauses facilitated the learning of dependency relations in both moved-syllable and novel-syllable conditions when analysing the combined data. We suggest that the novel intervening syllables in the test items led to too much variability in performance to detect any effect when analysing this dataset alone. Another possibility is that the difference between the condition with pauses and the condition without pauses is larger in the experiment with moved intervening syllables than in the experiment with novel intervening syllables, due to the below-chance performance found in the moved-syllable condition after familiarisation without

pauses. Interestingly, in the novel-syllable condition, we do not find the same below-chance performance. The enhanced performance in Experiment 2 strengthens the idea that the part-words encountered in the familiarisation stream in Experiment 1 were harder to reject compared to rule-words of the form A_iAC_i or A_iCC_i because part-words were statistically more likely in their original form, and not because children were not able to generalise the NADs. When listening to a continuous stream without pauses, this resulted in below-chance performance in Experiment 1 (45.6%), which disappeared in Experiment 2, with overall results remaining at chance-level (50.78%).

Using novel intervening syllables in the test stimuli may inhibit participants' need to suppress any part-words that had occurred in the stream spanning word boundaries (as in Experiment 1), but it does not enhance learning in such a way that children no longer need other segmentation cues. This result suggests that while using novel intervening syllables in the test stimuli does prevent below-chance performance, due to the inability to reject part-words that occur in the familiarisation stream, it does not allow children to detect the underlying NADs. This is in contrast with the results found by Frost and Monaghan (2016). In their study with adults, using novel intervening syllables in the test stimuli yielded evidence for learning without any other cues segmenting the familiarisation stream.

In the current study, children only chose significantly more rule-words than part-words containing novel intervening syllables when the familiarisation stream was segmented by pauses. When the familiarisation stream contained pauses, we observed an increase to 53.6% correct scores in Experiment 2. This is significantly above chance-level (i.e., a score of 50%). In other words, school-age children do not benefit from novel intervening syllables in the way that adults do. In contrast to the findings by Soderstrom et al. (2007) for infants, we did not find a clear hindering effect of novel elements either, although children did seem to benefit less from the presence of pauses when being tested on novel intervening syllables. In the mixed-effects model, Pause did not significantly contribute beyond the random effects to explain differences in performance when testing children using novel intervening syllables. However, when comparing the proportion of correct responses against chance-level, we found that only the group of children who received the familiarisation stream with pauses performed above chance. These results suggest that school-age children are more likely to extract an artificial NAD rule during passive listening when an additional segmentation cue in the form of a pause is present in the speech stream, and they can then detect this rule across familiar and novel speech items.

The children in the current study obtained overall lower accuracy scores compared to the adults in Peña et al. (2002) and Frost and Monaghan (2016), with both studies reporting an accuracy rate of around 70% in the moved-syllable condition with pauses ($d = 1.307$) and in the novel-syllable condition without pauses ($d = 1.209$) respectively.

Nonetheless, we obtained an effect size of 0.294 in the condition with pauses added to the familiarisation stream, which can be considered a small but meaningful effect of pauses. The smaller effect size may be attributed to the fact that adults are better explicit learners than children (Ferman & Karni, 2010; Ojima & Okanoya, 2020). In addition, the children watched a silent animation movie while listening passively to the familiarisation stream, while the adults in Peña et al. (2002) and Frost and Monaghan (2016) did not perform any other task during the listening phase. We believe that for our age group, engaging in active listening to the stream for 10 minutes is not feasible. However, this may have hindered the active learning process. It should also be noted that even though we found a significant increase in scores when pauses were added to the stream, our data showed that not all children in our study were able to learn the NAD rules. We found a significant positive correlation between children's age and the number of correct items, indicating that older children show more evidence of learning NADs. This age-related increase in performance may be due to the learning mechanisms that are used by the children in this task, reinforcing the idea that explicit learning performance increases with age. Older children may have more explicit metalinguistic knowledge (see Bialystok, 1986), which may have been beneficial in the present task. This supports previous findings by, for example, Ferman and Karni (2010) and Ojima and Okanoya (2020), who found more rule-learning success among adults and older children in artificial grammar learning paradigms when using a task that requires more explicit knowledge. It is also important to note that older children may simply have been better at understanding the task at hand. Even though there is no evidence to believe otherwise, we have not explicitly tested whether the children in our study understood the task, for example, by giving them the same task with familiar natural stimuli. It would be useful to do this in future studies, to be able to more reliably conclude that older children were better at learning NADs.

Conclusion

The results of this study strongly suggest that 7- to 11-year-old children have a better chance at learning artificial NADs when pauses are present in the speech stream. These pauses help divide the continuous stream into smaller chunks, making it easier to detect regularities within those chunks. The results also show that older children were more successful at detecting NADs than younger children, but that, overall, the children in our study were not as successful as the adults in previous studies based on the results of a forced-choice task. Nevertheless, performance was enhanced when prosodic cues, in the form of pauses, were added to the familiarisation stream. Only then were children able to discover the underlying NAD-rules across test items containing both moved and novel intervening elements. This reinforces the idea that prosodic cues facilitate language learners of all ages in discovering grammatical rules.

References

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear Mixed-Effects Models Using S4 Classes (R Package Version 0.999999-0).

Bialystok, E. (1986). Factors in the growth of linguistic awareness. *Child Development, 57*(2), 498–510. https://doi.org/10.2307/1130604

Boersma, P. & Weenink, D. (2018). Praat: doing phonetics by computer. Version 6.0.37. Retrieved from http://www.praat.org/.

de Diego-Balaguer, R., Rodríguez-Fornells, A., & Bachoud-Lévi, A.-C. (2015). Prosodic cues enhance rule learning by changing speech segmentation mechanisms. *Frontiers in Psychology, 6,* 1478. https://doi.org/10.3389/fpsyg.2015.01478

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition, 105*(2), 247–299. https://doi.org/10.1016/j.cognition.2006.09.010

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language, 60*(3), 351–367. https://doi.org/10.1016/j.jml.2008.10.003

Ferman S., & Karni, A. (2010). No childhood advantage in the acquisition of skill in using an artificial language Rule. *PLoS One 5*(10): e13648. https://doi.org/10.1371/journal.pone.0013648

Frost, R. L. A., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition, 147*, 70–74. https://doi.org/10.1016/j.cognition.2015.11.010

Grama, I. C., Kerkhoff, A., & Wijnen, F. (2016). Gleaning structure from sound: The role of prosodic contrast in learning non-adjacent Dependencies. *Journal of Psycholinguistic Research, 45*(6), 1427–1449. https://doi.org/10.1007/s10936-016-9412-8

Grama I. C., & Wijnen, F. (2018). Learning and generalizing non-adjacent dependencies in 18-month-olds: A mechanism for language acquisition?. *PLoS* One *13*(10), e0204481. https://doi.org/10.1371/journal.pone.0204481

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package:

Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(1), 1–26. https://doi.org/10.18637/jss.v082.i13

Lammertink, I., Witteloostuijn, M. V., Boersma, P., Wijnen, F., & Rispens, J. (2019). Auditory statistical learning in children: Novel insights from an online measure. *Applied Psycholinguistics, 40*(2), 279–302. https://doi.org/10.1017/S0142716418000577

Long, J.A. (2020). jtools: Analysis and Presentation of Social Scientific Data. R package version 2.1. https://cran.r-project.org/package=jtools

Marchetto, E., & Bonatti, L. L. (2013). Words and possible words in early language acquisition. *Cognitive Psychology*, 67(3), 130-150. https://doi.org/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. https://doi.org/10.1017/S0305000914000452

Marimon, M., Hofmann, A., Veríssimo, J., Männel, C., Friederici, A. D., Höhle, B., & Wartenburger, I. (2021). Children's learning of non-adjacent dependencies using a web-based computer game setting. *Frontiers in Psychology*, 12, 734877. https://doi.org/10.3389/fpsyg.2021.734877

Morgan, J. L. (1986). *From Simple Input to Complex Grammar*. MIT Press. https://mitpress.mit.edu/books/simple-input-complex-grammar

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*(5), 892–905. https://doi.org/10.1162/jocn.2008.20511

Mueller, J. L., Friederici, A. D., & Männel, C. (2012). Auditory perception at the root of language learning. *Proceedings of the National Academy of Sciences, 109*(39), 15953–15958. https://doi.org/10.1073/pnas.1204319109

Mueller, J. L., Milne, A., & Männel, C. (2018). Non-adjacent auditory sequence learning across development and primate species. *Current Opinion in Behavioral Sciences, 21*, 112–119. https://doi.org/10.1016/j.cobeha.2018.04.002

Ojima, S. & Okanoya, K. (2020). Children's learning of a semantics-free artificial grammar with center embedding. *Biolinguistics, 14*, 21-48.

Pacton, S., & Perruchet, P. (2008). An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 34*(1), 80–96. https://doi.org/10.1037/0278-7393.34.1.80

Paul, M., Männel, C., van der Kant, A., Mueller, J. L., Höhle, B., Wartenburger, I., & Friederici, A. D. (2021). Gradual development of non-adjacent dependency learning during early childhood. *Developmental Cognitive Neuroscience, 50*, 100975. https://doi.org/10.1016/j.dcn.2021.100975

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298*(5593), 604–607. https://doi.org/10.1126/science. 1072901
Perruchet, P., Tyler, M. D., Galland, N., & Peereman, R. (2004). Learning nonadjacent dependencies: No need for algebraic-like computations. *Journal of Experimental Psychology. General, 133*(4), 573–583. https://doi.org/10.1037/0096-3445.133.4.573

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*(4), 606–621. https://doi.org/10.1006/jmla.1996.0032

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods, 8*(4), 448-467.

Schaadt, G., Paul, M., Muralikrishnan, R., Männel, C., & Friederici, A. D. (2020). Seven-year-olds recall non-adjacent dependencies after overnight retention. *Neurobiology of Learning and Memory, 171*, 107225. https://doi.org/10.1016/j.nlm.2020.107225

Skeide, M. A., & Friederici, A. D. (2016). The ontogeny of the cortical language network. *Nature Reviews Neuroscience, 17*(5), 323–332. https://doi.org/10.1038/nrn.2016.23

Soderstrom, M., Seidl, A., Kemler Nelson, D. G., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language, 49*(2), 249–267. https://doi.org/10.1016/S0749-596X(03)00024-X

Soderstrom, M., White, K.S., Conwell, E. and Morgan, J.L. (2007). Receptive grammatical knowledge of familiar content words and inflection in 16-month-olds. *Infancy*, *12*(1), 1-29. https://doi.org/10.1111/j.1532-7078.2007.tb00231.x

van der Kant, A., Männel, C., Paul, M., Friederici, A. D., Höhle, B., & Wartenburger, I. (2020). Linguistic and non-linguistic non-adjacent dependency learning in early development. *Developmental Cognitive Neuroscience, 45*, 100819. https://doi.org/10.1016/j.dcn.2020.100819

Zhang, Z., & Yuan, K. H. (2018). *Practical Statistical Power Analysis Using Webpower and R* (Eds). Granger: ISDSA Press.

Data, code and materials availability statement

Materials, data, and scripts are available online: https://osf.io/e43hw/

Ethics statement

The research was conducted in accord with the Research Ethics Code of the Leiden University Centre for Linguistics. The parents of all the children participating in the study signed an informed consent form that was sent to them through the schools that participated in the research project.

Authorship and Contributorship Statement

Van der Klis and Van Lieburg designed the experiment, collected the data, performed the statistical analyses, and drafted and revised the manuscript. Cheng and Levelt were responsible for the conception of the study, edited the manuscript and supervised the study. All authors are responsible for the interpretation of the results. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

This work was supported by the Research Traineeship Programme of the Faculty of Humanities, Leiden University. We would like to thank the children for their participation, and Daltonschool De Margriet in Rotterdam and Haanstra primary school in Leiden for their cooperation. We thank Anouk Vinders for her help at the Daltonschool De Margriet. We also want to express our gratitude to Dr. Leticia Pablos Robles for offering us extensive statistical advice.

License

Some puzzling findings regarding the acquisition of verbs

Joshua K. Hartshorne
Yujing Huang
Lauren Skorb
Boston College, USA

Abstract: On the whole, children acquire frequent words earlier than less frequent words. However, there are other factors at play, such as an early "noun bias" (relative to input frequency, toddlers learn nouns faster than verbs) and a "content-word bias" (content words are acquired disproportionately to function words). This paper follows up reports of a puzzling phenomenon within verb-learning, where there appears to be a large effect of argument structure class, such that verbs in one class (experiencer-object verbs) were learned substantially earlier than those in another (experiencer-subject verbs) despite being much lower frequency. In addition to the possibility that the aforementioned results are a fluke or due to some confound, prior work has suggested several possible explanations: experiencer-object ("frighten-type") verbs have higher type frequency, encode a causal agent as the sentential subject, and perhaps describe a more salient perspective on the described event. In three experiments, we cast doubt on all three possible explanations. The first experiment replicates and extends the prior findings regarding emotion verbs, ruling out several possible confounds and concerns. The second and third experiments investigate acquisition of chase/flee verbs and give/get verbs, which reveal surprising findings that are not explained by the aforementioned hypotheses. We conclude that these findings indicate a significant hole in our theories of language learning, and that the path forward likely requires a great deal more empirical investigation of the order of acquisition of verbs.

Keywords: language acquisition, perspective pairs, psych verbs, verb-learning

Corresponding author: Joshua K. Hartshorne, Department of Psychology and Neuroscience, Boston College, 140 Commonwealth Ave., McGuinn 522, Cambridge, MA 02467. Email: joshua.hartshorne@bc.edu.

ORCID ID: https://orcid.org/0000-0003-1240-3598

Citation: Hartshorne, Joshua K., Huang, Y., & Skorb, L. (2023). Some puzzling findings regarding the acquisition of verbs. *Language Development Research, 3*(1), 65–104. https://doi.org/10.34842/2023.535

Introduction

Learning a word is more than just a function of having heard it. On the whole, children do acquire frequent words earlier than less frequent words, but that is hardly the end of the story (Goodman, Dale, & Li, 2008; Hansen, 2017). Controlling for frequency, nouns are learned earlier than verbs, which are in turn learned earlier than closed-class words (Gentner, 1982; Goodman et al., 2008; Hansen, 2017). Controlling for word type, highly-imageable words are learned earlier than less-imageable words, perhaps because it is easier to identify the intended referent during conversation (Hansen, 2017; McDonough, Song, Hirsh-Pasek, Golinkoff, & Lannon, 2011). Indeed, raw frequency may be the wrong measure: rather than word knowledge slowly accumulating with each exposure, word-learning may be disproportionately driven by highly-informative learning opportunities, which give rise to "eureka" moments (Medina, Snedeker, Trueswell, & Gleitman, 2011).

One issue particular to learning verbs is learning how they convey who does what to whom. A child who knows no more about *bite* than what sort of action it describes will be at a loss to distinguish between *dog bites man* and *man bites dog*. To really master a verb, she must know its argument structure: which event roles (biter, bite-ee) are realized in what syntactic positions (subject, direct object). The need to learn argument structure might partly explain why verbs are acquired more slowly than nouns.

Recent findings from Hartshorne, Pogue, & Snedeker (2015) suggested that argument structure may also explain why certain verbs are harder to learn than others. In particular, they found that "psych verbs" that realize the experiencer as the direct object (*A frightened/pleased/angered B*; "frighten-type verbs) are acquired substantially earlier than those that realize the experiencer as the subject (*A feared/liked/hated B*; fear-type verbs), despite being much lower frequency. Specifically, while English-speaking children have already acquired a handful of frighten-type verbs by the age of 4, they do not start acquiring fear-type verbs until about a year later. In particular, although four year-olds use words like *like* and *love* in spontaneous speech, they struggle to distinguish who did what to whom, treating *A loves B* as equivalent to *B loves A* (and similarly for other fear-type verbs). Illustrating just how unexpected this finding was, it had actually been observed in prior studies but never taken seriously, being instead either dismissed as the result of confounds or not remarked upon at all (Bowerman, 1990; Braine, Brooks, Cowan, Samuels, & Tamis-LeMonda, 1993; Messenger, Branigan, McLean, & Sorace, 2012; Tinker, Beckwith, & Dougherty, 1988); by addressing those confounds and reporting multiple targeted studies using different methods, Hartshorne, Pogue, and Snedeker (2015) established the finding as something in need of explanation.

Hartshorne and colleagues interpret this finding as evidence for a "**privileged link**" between causality and sentential subjects. Specifically, a long line of research suggests both a cross-linguistic tendency for the agents of caused events to be mapped onto

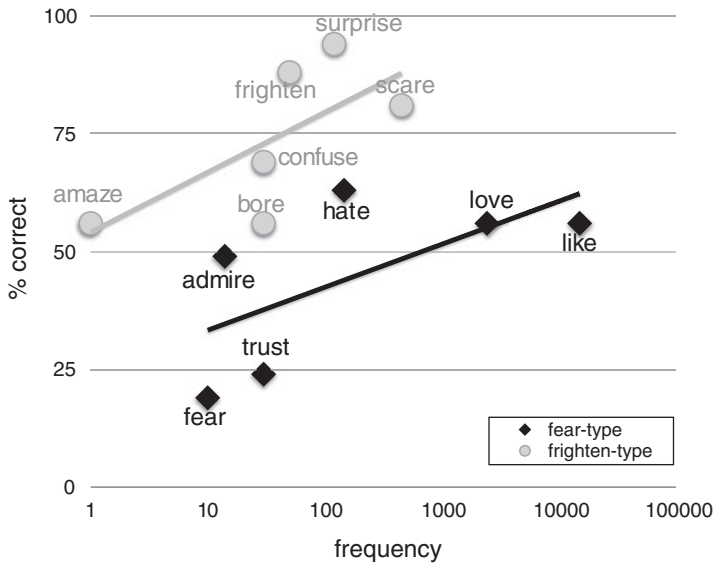


Figure 1. ***Taking frequency in child-directed speech into account, four year-olds were substantially more likely to successfully interpret who did what to whom for frighten-type verbs relative to fear-type verbs (Hartshorne et al., 2015). Figure used with permission.***

subject position and a corresponding learning bias on the part of young children to expect agents of caused events to be subjects of corresponding sentences (Braine, 1992; Dowty, 1991; Fisher, Gertner, Scott, & Yuan, 2010; Levin & Hovav, 2005; Lidz, Gleitman, & Gleitman, 2003; MacWhinney, 1977; Marantz, 1982, 1982; Pinker, 1984; Strickland, Fisher, Keil, & Knobe, 2014). This set of findings has prompted a number of theorists – both Empiricist and Nativist – to argue for an innate bias for agents of caused events to be realized as sentential subjects (or whatever the analog of SUBJECT is in the account adopted by the theorist), and that this bias is a key part of what makes langauge learnable in the first place (for discussion, see Hartshorne et al., 2016). Critically, across a variety of languages, frighten-type verbs do encode causality (*A frightened B* means, roughly, "A caused B to feel fear") whereas fear-type verbs do not (*A feared B* means, roughly, "A was disposed to feel fear about B") (Hartshorne et al., 2016). (Note that while these findings overturned some earlier proposals that treated *fear* and *frighten* as synonymous, differing only in the syntax (Belletti & Rizzi, 1988; Dowty, 1989; Grimshaw, 1990), those had never been tested empirically.) Thus, the "privileged link" would give children a leg up in learning frighten-type verbs, and might even impede the acquisition of fear-type verbs.

The data reviewed above are certainly consistent with the privileged link hypothesis, but you can draw an infinite number of lines through a single point; there are a number of reasonable alternative explanations. One alternative explanation is **type frequency**:

While fear-type verbs have high token-frequency, they are relatively rare as types. English admits around 251 verbs that use frighten-type argument structure but only 49 fear-type verbs (Kipper, Korhonen, Ryant, & Palmer, 2008). The difference becomes even more stark when one considers that frighten-type verbs are a special case of an extremely robust pattern (caused changes of state being realized in transitive syntax with the AGENT as the subject and the PATIENT as the direct object), whereas fear-type verbs are quite unusual (only a handful of other verbs have EXPERIENCERs as subjects). According to a variety of learning theories – including both Nativist and Empiricist – this high type frequency should benefit frighten-type verbs, for instance by helping learners identify the set of features of verbs that reliably predict which verbs admit that particular argument structure (Goldberg, 2006; Pinker, 1989).

A second alternative is **differential salience**: Most utterances are contextually ambiguous: there are many things someone could be talking about (Gleitman, 1990; Quine, 1964). For instance, it is often the case that when there is fearing going on, there is also some frightening; the speaker can choose which to comment on. If children find frightening more salient than fearing, they may be more disposed to successfully identify labeling of frightening than labeling of fearing. Unfortunately, no data currently speak to whether there is a salience asymmetry between frighten-type events and fear-type events.

There are a number of other, less easily testable options. For instance, it is possible that children's exposures to fear-type verbs are relatively short on highly informative "eureka" moments, making them effectively lower-frequency. While there are methods for quantifying the availability of eureka moments, they are time-consuming and difficult to scale (cf. Medina et al., 2011). It may be that the thoughts encoded by fear-type verbs (a habitual disposition towards a particular emotion) are more complex, harder to represent, or emerge later in cognitive development than those encoded by frighten-type verbs (an externally-caused change of emotional state). It is not clear at the moment how this possibility would be tested. It may be that acquisition of frighten-type verbs is aided by the acquisition of other, related argument realization patterns, or that the acquisition of fear-type verbs is impeded by interference from other, contrasting argument realization patterns. This is currently difficult to test because the nature of argument realization patterns remains highly controversial (Levin & Hovav, 2005), and because very little is known about their acquisition outside of a handful of patterns and languages (for comprehensive reviews, see Ambridge et al., 2018, 2020).

There are no doubt other possibilities as well, including of course the possibility that the psych verb findings are a statistical fluke: It happens that children learn their first few frighten-type verbs before learning their first few fear-type verbs, but the reasons are idiosyncratic to each verb and have nothing in particular to do with argument structure class.

Goals of the present study

The present study has two main goals: a) to begin to assess the robustness and generalizability of the finding that argument structure class modulates the timeline of verb acquisition, and b) if the finding is robust and generalizable, tease apart the three hypotheses highlighted above.

A challenge in assessing generalizability is that there are hundreds of argument structure classes in English alone (groups of verbs that use distinct argument structures) (Kipper et al., 2008). Collecting data on even just a substantial portion of them is a long-term project. Unfortunately, we cannot lean much on pre-existing data, since there is almost no prior work identifying the ages of acquisition of the argument structures of specific verbs. Studies of argument structure knowledge typically involve older children who already know quite a few verbs in that class (Ambridge et al., 2018; Pinker, 1989). Studies of vocabulary emergence largely depend on spontaneous production by the child or parental report that the child knows the word, neither of which directly assesses knowledge of argument structure and may in fact be uncorrelated with such knowledge (Hartshorne et al., 2015).

Thus, in order to increase the informativity of the current project, we focused on "perspective pairs". These are groups of verbs which describe similar types of events but contrast in which event-participant is realized as the sentential subject. Psych verbs are an example, where *A feared B* and *B frightened A* can both be said of the same event. While not every fear-type verb has a frighten-type counterpart, the classes as a whole systematically differ in argument structure while describing highly similar types of events.

Focusing on perspective pair classes has the distinct advantage of minimizing uncontrolled differences across verbs. Thus, perhaps children learn *frighten* later than *kick* because the former involves describing a mental state, something that young children struggle with (Wellman, 1992). Or perhaps it is due to any of the myriad other ways that frightening differs from kicking. In contrast, there are far fewer semantic differences between frighten-type and fear-type verbs, and the ones that exist are reasonably well understood (such as how they encode causation).

A second advantage of perspective pairs is that they provide perhaps the best opportunity for testing the salience hypothesis described above. As our science develops, we may eventually have good mechanisms for quantitatively comparing the relative salience of events of frightening vs. events of kicking, but currently this is quite difficult. This question is more straightforward for perspective pairs, and indeed there is some prior work that can inform our investigation (see review below).

We conducted three experiments. The first replicates and extends Hartshorne, Pogue, &

Snedeker (2015) with a larger number of psych verbs (16 vs. 12) and a wider range of ages (3-6 vs. 4-5), while simultaneously allowing us to address some possible concerns about that study's methods. If the prior findings do not generalize at least this much, that would fundamentally change the question. Experiments 2 and 3 focused on chase/flee verbs and give/get verbs, respectively. Chase-type and flee-type verbs differ in terms of whether the pursuer is the subject (*A chased/pursued/followed B*) or an oblique object (*A fled/escaped/ran from B*).[1] Give-type and get-type verbs differ in whether the SOURCE is the subject and the GOAL is an oblique (*A gave/passed/sold B to C*) or *vice versa* (*A got/grabbed/bought B from C*). (Note that, like nearly all verbs, fear/frighten, chase/flee and give/get verbs can appear in other sentence frames as well; here, we focus on the ones that show the contrast most cleanly.)

Critically, these three case studies set up a substantial number of clear comparisons with respect to token frequency, causality, type frequency, and salience (Table 1). The frequency comparisons are straightforward and presented in Tables 2, 4, & 6). We discuss the other three comparisons in detail below.

Table 1: For each of several factors, if that factor was determinative, which verbs would be acquired first.

	Exp1	Exp2	Exp3
pair	fear/frighten	chase/flee	give/get
token frequency	fear	chase	neither
causality	frighten	chase	give
type frequency	frighten	chase	give
salience	??	?chase	get

Causal Semantics

As reviewed above, the "privileged link" for causality hypothesis predicts earlier learning for frighten-type (which have a causal semantics) relative to fear-type (which do not). However, while semantic analysis typically ascribes causality to the subject of both give-type and get-type verbs, and to *neither* chase-type nor flee-type (Kipper et al., 2008; Pinker, 1989), there do not appear to have been any systematic studies (Hartshorne, Bonial, & Palmer, 2014).

[1]While chase-type and flee-type verbs can describe the same events, there is evidence that they – like fear-type and frighten-type verbs – are semantically distinct. Gleitman (1990) reports a personal communication from Steven Pinker arguing that intentional participation in the event is entailed only for the subject of chase/flee verbs: one can chase something that is not fleeing (e.g., a storm) or flee something that is not chasing (e.g., a tsunami). So far as we know, there has not been any systematic study of the verb classes to determine whether this generalizes, though initial inspection suggests that it does. In any case, the exact semantics of these verbs will not be critical for the present study.

We conducted a study in order to obtain quantitative, empirical measurements.[2] The task involved judging who caused various events to happen: specifically, 30 events involving each of the 30 verbs used across the three experiments below (see Tables 2, 4, & 6). Thus, participants were asked to answer questions of the form:

> Who made this happen?: Agnes frightened Beatrice.
>
> Ages, Beatrice, Both of them, Someone else, Nobody (these things just happen), Can't tell

Note that the primary response of interest was how many participants assigned causality to the sentential subject; the other options were included in order to provide natural alternatives.[3]

We recruited native English-speaking adults through Prolific, aiming for 100 participants after exclusions. The final sample was 101 (mean age = 39, range = 18 - 68). An additional 53 participants were excluded for missing one or more catch trials where the answer to the question was stated explicitly (ex: "Who made this happen?: Agnes made Beatrice do something."). There were five catch trials targeting the five primary judgments of interest (all except "can't tell"). This ensured that participants who were included understood roughly what judgments we wished them to make.

As expected, participants were far more likely to judge the subject of frighten-type verbs to have caused the event (88%) than subjects of fear-type verbs (44%) (Table 6). In contrast to prior linguistic analyses, participants also judged the subject of chase-type verbs to be more likely to cause the event (92%) than subjects of flee-type verbs (63%) (Table 4), and the subject of give-type verbs (91%) more than the subjects of get-type verbs (46%) (Table 6). For each of the perspective pairs, the differences between classes were categorical, with the exception of unusually large variability across the four get-type verbs.

Thus if the "privileged link" explains the early learning of frighten-type verbs, we should also expect earlier learning of chase-type vs. flee-type verbs and possibly give-type vs. get-type verbs.

Type Frequency

We assessed type frequency using VerbNet, the largest compendium of verb classes in English (Kipper et al., 2008). In terms of type-frequency, as reviewed above, frighten-

[2]We thank an anonymous reviewer for this suggestion

[3]We did inspect the rates at which participants chose the other answers in order to confirm that the results were sensible, but we did not systematically analyze them, other than the analysis described in footnote 13.

type verbs (N = 251) are far more numerous than fear-type (N = 49).

Chase-type verbs (N = 22) are more common than flee-type, of which there appear to only be around 4. Counting is complicated in that while VerbNet records three groups of chase-type verbs (classes 35.1, 35.3, and 51.6), it does not index flee-type verbs, perhaps because there are so few. *Flee from*, *escape from*, and *retreat from* are included in class 51.1, and *run from* in class 51.3.2, but only in the sense of escaping from a *place,* not an entity. Because flee-type verbs are not indexed in VerbNet, it is possible there are more that we have overlooked, though probably not enough to place the class on par with chase-type. Interestingly, there is a class of verbs (class 52; N = 11) in which the fleer is the subject and the pursuer is the direct object (*A dodged/eluded/evaded B*). Because they take direct objects, they belong in a different syntactic category from chase/flee verbs. We did not investigate them in addition to or instead of chase/flee verbs because they are all low-frequency and unlikely to be encountered by children.

Finally, give-type verbs (N = 82; classes 13.1, 13.2, 13.4.1, 13.4.2) are more numerous than get-type verbs (N = 57; classes 13.5.1, 13.5.2, 57), but the imbalance is not as stark as for the other perspective pair classes.[4]

Salience

As already noted, there is no evidence bearing on the question of whether frighten-type event construals are more or less salient than fear-type construals, and we do not investigate it here (it is currently clear how to do so). With regards to chase/flee, it has been argued that children and adults are biased to encode ambiguous events in terms of chasing rather than fleeing (e.g., Landau & Gleitman, 2015). However, the evidence is thin and mixed. Fisher, Hall, Rakowitz, and Gleitman (1994) indeed found that three year-olds and adults described one ambiguous chase/flee scene in terms of chasing rather than fleeing. However, Gleitman, January, Nappa, and Trueswell (2007) ran a similar study with adults using two stimuli, finding a chase-bias for one and a flee-bias for the other. There do not appear to be any other empirical studies.

There is a more robust literature indicating that when presented with an ambiguous give/get event, both children and adults focus on the getting over the giving. Children and adults are more likely to remark on and remember GOALS than SOURCES (Freeman, Sinha, & Stedmon, 1981; Fujita, 2000; Lakusta & Landau, 2005, 2012; Papafragou, 2010; Regier & Zheng, 2007). This suggests that get-type verbs should be privileged for two reasons. First, the GOAL is obligatory for get-type verbs and optional for many give-type verbs (*A sent/sold/passed B [to C]*). Second, it is widely argued that more salient entities

[4]We exclude verbs of future having (*A awarded/bequeathed/owed B to C*) from give-type verbs. The semantics are critically different in that there is no caused change of possession, just a promised change of possession. If these are included in the count, the advantage in token-frequency for give-type verbs becomes more imbalanced at 110 to 57.

are preferentially mapped onto sentential subjects.

In terms of language acquisition, there is only limited evidence as to an advantage for get-type (or give-type) verbs. In studies looking at spontaneous description, participants often avoid both give-type and get-type verbs, instead preferring THEME-subject verbs (*A walked/ran/rolled to B*). Lakusta and Landau (2005) similarly found that children were more likely to use known give-type verbs than get-type verbs when describing ambiguous scenes (Lakusta & Landau, 2005).[5] A smaller novel-word learning study showed a similar bias towards give-type verbs (Fisher et al., 1994). However, Bowerman (1990) reports in a diary study of two children that the first get-type verb (*get*) emerges in spontaneous speech only just prior (1;8) to the first give-type verbs (*give/gimme*, *tell*, and *read me*, all of which emerge at 1;9). Since get-type verbs appear first but give-type appear in larger numbers, this seems to be a draw.

Summary of Predictions

The predictions for the three experiments are summarized in Table 1.

- If token frequency is the key factor, we would expect earlier acquisition of fear-type verbs compared to frighten-type; chase-type compared to flee-type; and roughly equal learning for give-type and get-type. This outcome seems unlikely given the results of Hartshorne, Pogue, and Snedeker (2015), but perhaps our replication-and-extension will show different results.
- If the "privileged link"/causality hypothesis is correct, we should see earlier acquisition of frighten- type than fear-type verbs, chase-type verbs than flee-type, and give-type relative to get-type.
- If type frequency is the key factor, we should see earlier acquisition of frighten-type than fear-type verbs; chase-type relative to flee-type; and give-type relative to get-type.
- If the salience hypothesis is correct, we have no strong predictions for psych verbs, but have a weak prediction of earlier learning of chase-type than flee-type verbs, and a stronger prediction of earlier learning of get-type than give-type.

By using these three case studies, we hope to disentangle the four main hypotheses under consideration.

[5] Lakusta and Landau (2005) argue that the salience hypothesis predicts a preference for give-type verbs. In particular, they posit that children should prioritize the mapping to the prepositional phrase and thus find it easier to map the more salient GOAL onto the prepositional phrase, rather than the SOURCE. This is an intriguing notion, but it is incompatible with most prominent theories of language acquisition, and it remains to be seen whether a new theory can be constructed around it. It is also inconsistent with the data we present below. Thus, we do not consider this hypothesis further in this paper.

Overview of analyses

In all three experiments, we submit the data to a mixed effects logistic regression with centered main effects of verb type, log frequency, and subject age in months.[6] In order to improve convergence and avoid issues with singularity, we fit the model with partially-Bayesian regression with Wishart priors on the covariance matrix for random effects, using the blme package with bobyqa optimization (Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013).

We included random intercepts by subject and verb as well as a random slope of verb type by subject. We chose not to use a maximal random effects structure for three reasons. First, there is some debate about whether doing so is even desirable (Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). In particular, specifying overly complex models may substantially increase Type II error. Indeed, we found that for several of our analyses, a maximal random effects structure resulted in *no* significant effects, not even effects of age or log frequency, even though these are extremely well attested in the literature and clearly visible in our data. Second, the small number of verbs (ranging from 6 to 16 per study) means that random slopes are estimated based on very few data points. For instance, estimating a random slope for participants for the interaction of verb type and log frequency involves calculations of slopes based on as few as 3 and at most 8 boolean responses. Third, empirically we found that in most cases random slopes led to worse fits (as assessed by BIC and AIC) – when fitting could even be done successfully. Analysis of the random effects structure for Exp. 3 did justify random slopes of verb type and log frequency by subject, but we felt using a different random effects structure for different experiments would impede comparison across experiments. Thus, we settled on a relatively simple random effects structure with only one random slope (verb type by subject), which is a fairly simple slope based on a relatively large amount of data.

While this is our primary analysis, we considered two other analysis methods. First, we fit the primary models using Bayesian regression as implemented in brms (Bürkner, 2017, 2018) and calculated Bayesian p-values with the p_map function from bayestestR (Makowski, Ben-Shachar, & Lüdecke, 2019).[7] Unless data sets are very large, Bayesian analyses tend to be biased in favor of the null hypothesis, so it is not surprising that some of the effects that were significant in our main analyses are not significant in the Bayesian regressions, though for the most part the key results remained the same. The results of these models are reported in footnotes (in general, we use footnotes in

[6] Not centering these variables frequently led to expected failures to converge.

[7] For the intercept, fixed effects, and random effects, we used the relatively informative prior of a half-normal distribution centered at 0 and with a standard deviation of 1. We ran six chains of 8000 samples each, including 1000 warmup samples. For two of the models, we raised alpha_adapt to 0.9 in order to address a small number of divergent transitions.

lieu of supplementary materials). Second, we used model comparison to prune fixed effects from the primary model. Unfortunately, we obtained wildly different results using AIC, which generally favored retaining most or all of the fixed effects, and BIC, which is a very conservative method – particularly when data sets are not very large – and which generally favored eliminating most or even all of the fixed effects. Because the effects of log frequency and age are well-established and quite clear in the data, this suggests BIC is overly conservative. We do not describe these results in detail, but they are memorialized in our reproducible RMarkdown document, available at osf.io/k5xud. Because the Bayesian regression and BIC method both appear to be overly conservative, we do not consider their results as strongly indicative of the null, but it certainly does mean that the evidence we present below is not incontrovertibly strong.

We also report follow-up analyses for each experiment, focused on each individual age group (3 year-olds, 4 year-olds, 5 year-olds, and 6 year-olds).

R packages used to prepare this reproducible document include *papaja* (Aust & Barth, 2020), *knitr* (Xie, 2015), *ggplot2* (Wickham, 2016), *ggeffects* (Lüdecke, 2018), and *stargazer* (Hlavac, 2018).

Experiment 1: Fear/Frighten (Psych) Verbs

This experiment replicated the method of Exp. 1 of Hartshorne, Pogue, and Snedeker (2015), with the following changes: a) we added two new fear-type verbs and two new frighten-type verbs, b) we expanded the age range to 3 to 6, c) in order to accommodate the attention spans of the younger participants, we split the verb lists so that each subject saw only half of the stimuli, and d) we presented half the participants with critical verbs in the past tense, half in present tense.

The manipulation of tense requires some explanation. Tense affects fear-type and frighten-type verbs differently. While fear-type verbs refer to a habitual state whether used the present tense (*A fears B* [*always / *just the once*]) or past tense (*A feared B* [*always / * just the once*]), frighten-type verbs most naturally refer to a habitual or repeated event in the present tense (*A frightens B* [*always / *just the once*]) and a single event in the past tense (*A frightened B* [*?always / just the once*]). Unfortunately, Hartshorne, Pogue, and Snedeker (2015) did not explicitly control the tense used, so it is unclear whether the same tense was used for all verbs. It is unlikely this could explain poorer performance on fear-type verbs, which are unaffected by tense, but out of an abundance of caution, Exp. 1 included a tense manipulation in order to test the question directly.

Method

All research reported in this paper was approved by the Boston College and Harvard University Institutional Review Boards.

Participants

We recruited 290 native English-speaking children between the age of 3 and 6 years from the Greater Boston Area from parks, museums, and preschools. Of these, we excluded 47 children due to coding error or failure to complete more than 50% of the experiment or both. We had intended to have 64 participants per age group (one per list; see below), but delays due to the pandemic made it impossible to reach the full complement in some cases. In others, we ended up with more participants than intended for the reasons described in "Procedure". The final numbers included 54 3 year-olds, 73 4 year-olds, 82 5 year-olds, and 34 6 year-olds.

Materials

A total of 16 verbs were tested (8 fear-type and 8 frighten-type), including the 12 verbs from Hartshorne, Pogue, and Snedeker (2015) and 4 additional verbs. Part of the goal was to include three relatively high-frequency verbs (*enjoy*, *dislike*, *bug*) that were overlooked in construction of the earlier experiments (Hartshorne, Pogue, and Snedeker (2015) did not have access to as complete a list of fear/frighten verbs as we do). The final verb (*anger*) is relatively low-frequency but was commonly used by children during a pilot verb-elicitation study and so was included.

We estimated frequencies for fear/frighten verbs in speech directed to children ages 36 to 84 months (the age range in the present studies) by using childesr (Sanchez et al., 2019) to aggregate all speech (other than that by the target child) in North American English corpora in CHILDES (MacWhinney, 2000) where the target child was in the age range and where part-of-speech tags were available. This aggregate corpus consisted of 3,044,358 tokens. Frequencies for our stimuli are shown in Table 2. A disproportionate number of high-frequency words were fear-type, though the difference between types across the 16 verbs did not reach significance ($\Delta M = 0.19$, 95% CI $[-2.14, 2.52]$, $t(11.44) = 0.18$, $p = .863$).[8]

Table 2: Verbs used in Experiment 1, with frequency in parts per million and probability that causality is assigned to the sentential subject.

Verb	Type	Frequency	LogFrequency	SubjCause
dislike	fear	0.3	0.3	50
admire	fear	1.6	1.0	48
fear	fear	2.3	1.2	27
trust	fear	4.6	1.7	45
enjoy	fear	48.9	3.9	40
hate	fear	55.2	4.0	48

[8]For these and other statistics, we used log-transformed frequencies.

Verb	Type	Frequency	LogFrequency	SubjCause
love	fear	269.0	5.6	47
like	fear	2246.5	7.7	48
anger	frighten	0.0	0.0	86
bug	frighten	8.9	2.3	87
frighten	frighten	15.8	2.8	90
confuse	frighten	16.4	2.9	83
amaze	frighten	20.0	3.0	94
bore	frighten	27.9	3.4	78
surprise	frighten	56.5	4.1	95
scare	frighten	214.2	5.4	94

For each of the 16 verbs, we created four scenarios, counterbalancing the pair of characters involved and which member of the pair experienced the emotion. We created two stimulus orders as follows: We did a mid-line split on token frequency for both fear-type and frighten-type verbs and placed the 4 highest-frequency of each type in the first half the list. We then created a second list reversing the orders *within* each half, so that the high-frequency verbs remained in the first half. The purpose of this was to allow participants to continue on to the second half if they have sufficient interest, but this rarely happened. Ultimately, we began randomly assigning participants to one half or the other. (For those participants from the initial phase of testing who had completed both halves, we excluded the second half.) The result was 64 lists: 2 (item set) by 2 (present vs past) by 2 (animal pair) by 2 (animal roles) by 2 (item order) by 2 (target response) design.

During testing, we discovered an error in the lists such that for four of the low-frequency verbs (*bug*, *anger*, *enjoy*, *dislike*) only one of the character pairs was used. We fixed the lists and began replacing those participants (N=27) who had not seen the intended character pairs. However, because we were unable to finish replacing them due to the pandemic, we included both the original and "replacement" participants in order to improve power. Nonetheless, excluding the to-be-replaced participants has no effect on the qualitative pattern of results.

Procedure

Participants were run one at a time. An experimenter read stories accompanied by pictures. After each story, a second experimenter used a Mickey Mouse puppet to say what Mickey thought happened in the story. Sometimes what Mickey said was correct and sometimes it was incorrect. If the participant thought that what Mickey said was correct, they gave Mickey a cookie. If they thought that what Mickey said was incorrect, they gave him coal. The experimenter began with two practice stories involving action

Table 3: Regression results for psych verbs

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.353	0.074	4.781	0.00000
Age	0.290	0.056	5.166	0.00000
VerbType	0.248	0.144	1.718	0.086
LogFrequency	0.246	0.082	3.020	0.003
Age:VerbType	0.380	0.108	3.530	0.0004
Age:LogFrequency	0.113	0.060	1.885	0.059
VerbType:LogFrequency	0.346	0.162	2.138	0.033
Age:VerbType:LogFrequency	0.196	0.119	1.649	0.099

verbs (*hug* and *kiss*), followed by the eight test trials. Participants received explicit feedback on the two practice trials but only general affirmative reactions to the critical trials. Responses were coded on site by a second experimenter. When possible, these were then double-checked from a video recording.

Results

Data for this and all experiments are available at osf.io/k5xud.

The dependent measure was accuracy: correctly accepting a true statement or rejecting a false one. We conducted a preliminary analysis to assess whether tense systematically affected the results. We submitted the results to a partially-Bayesian mixed effects logistic regression with Wishart priors on the covariance matrix for random effects, using the blme package with bobyqa optimization (Chung et al., 2013).[9] We included main effects of verb type and tense, along with their interaction, and random slopes of verb type by subject and tense by verb.[10] The main effect of tense was not significant (B = -0.13, SE = 0.15, Wald's z = -0.88, p=0.38), nor was the interaction of tense and type (B = 0.17, SE = 0.29, Wald's z = 0.59, p=0.56). Thus, all subsequent analyses collapsed across tense.

[9] In the interests of full communication of statistics, we provide more information rather than less. This includes providing exact p-values for all significant results, in order to support meta-analysis. For reasons of space, non-significant p-values are not reported except for marginal values (.05<p<0.10). While there is some debate over how to interpret marginal effects, we have erred on the side of providing information that readers may find useful. In any case, none of the marginal effects reported directly impinge on our conclusions one way or another.

[10] This is actually a maximal design. The random slope of verb type by subject is included in all analyses, so its inclusion requires no additional justification. The random slope of tense by verb is a fairly simple slope (the effect is categorical) and Exp. 1 has a relatively large number of verbs (16). In any case, excluding this random slope does not change the pattern of results. Note that the random slope of tense is not included in other analyses because it is not relevant to those analyses.

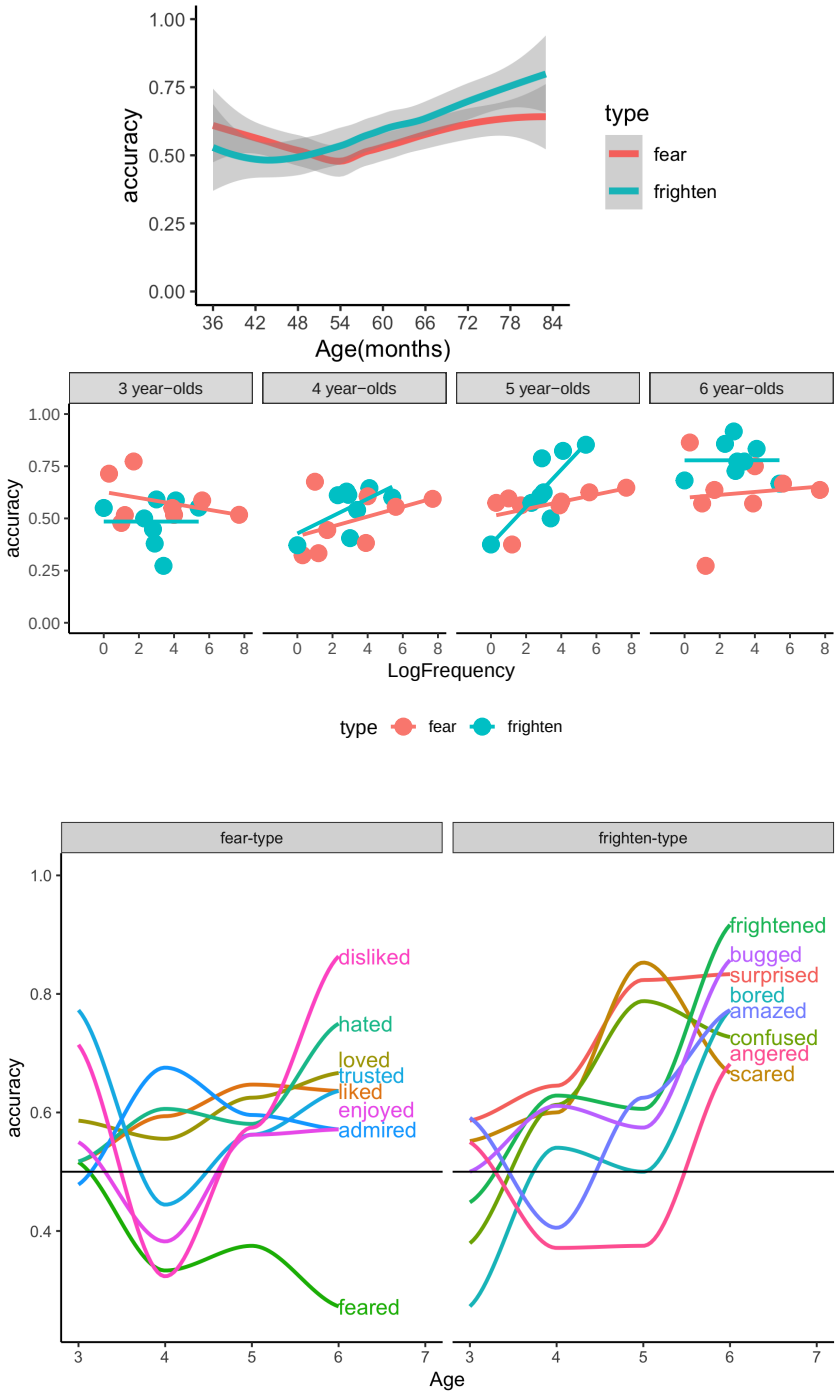


Figure 2. ***Results from Exp. 1. Top: Averaging across verbs within type, with LOESS smoothing over age. Middle: Accuracy for each verb against log frequency, split into four age groups, with linear regressions shown. Bottom: Performance on each verb across ages, aggregated into four age groups, with LOESS smoothing.***

We then conducted our main analyses, as described above in "Overview of Analyses." Results are shown in Table 3. The interaction of verb type by log frequency was significant, reflecting a larger effect of frequency on frighten-type verbs, as was the interaction of verb type by age, reflecting a larger advantage of frighten-type verbs in older participants.[11]

As explained in "Overview of Analyses," we followed up this analysis by considering each age group separately. For 3 year-olds, neither the main effects nor their interaction were significant (ps $\geq$ 0.12). For 4 year-olds, the only effect to approach significance was that of log frequency (B = 0.27, SE = 0.14, Wald's z = 1.90, p=0.057). By 5, there was a significant effect of log frequency (B = 0.49, SE = 0.13, Wald's z = 3.80, p=0.00014), a marginal effect of verb type (B = 0.41, SE = 0.23, Wald's z = 1.80, p=0.073), and a significant interaction (B = 0.71, SE = 0.25, Wald's z = 2.80, p=0.0055). By the age of 6, the only significant effect was that of verb type (B = 1.20, SE = 0.55, Wald's z = 2.10, p=0.033). Note that the sample sizes for 3 year-olds and 6 year-olds were smaller, perhaps explaining the fewer number of significant effects.

Fig. 2 visualizes these results along three dimensions. There is a clear overall advantage for frighten-type verbs beginning at around 50 months (Fig. 2 top). Interestingly, while there is a clear upward trajectory for frighten-type verbs beginning at around 48 months, there is no clear pattern for fear-type verbs, on which performance at 84 months is similar to that at 36 months.

Fig. 2 (middle) provides another window into the pattern of results: children exhibit higher performance on frighten-type verbs, controlling for frequency, by 4 years old, with the effect growing substantially by 6 years old. This figure is consistent with our age-group-specific analyses: 3 and 4 year-olds appear to be close to chance on most items, though there is a hint of better performance for high-frequency verbs by the age of 4. 5 year-olds show a clear pattern of success largely restricted to high-frequency frighten-type verbs. By the age of 6, children are doing well on all the frighten-type verbs but performance remains low for most fear-type verbs.

The apparent lack of improvement with age on fear-type verbs may be due to learners misclassifying low-frequency fear-type verbs as frighten-type (Fig. 2 Bottom). The highest-frequency fear-type verbs (*like*, *love*, *hate*) may not have reached the performance levels of frighten-type verbs, but they did show gradual improvement across the age range. In contrast, four of the lower-frequency verbs (*enjoy*, *trust*, *dislike*, *fear*) actually showed a *decline* from 3 to 4 years old, with all four verbs actually declining to below-chance levels. This matches the results of Hartshorne, Pogue, and Snedeker (2015), who reported significantly below-chance performance on *trust* and *fear* (*enjoy*

[11]The Bayesian regression revealed significant effects of age (p <0.00000001) and log frequency (p=0.034), and a significant interaction of age and verb type (p = 0.0023). The only result significant in the main analyses but not the Bayesian analyses was the interaction of frequency and verb type (p = 0.16)

and *dislike* were not tested).

Discussion

The results of Exp. 1 confirmed the results of Hartshorne, Pogue, and Snedeker (2015) with more verbs and across a larger age range: children began learning frighten-type verbs by the age of 4-5, whereas fear-type verbs remained largely at near-chance levels even at the age of 6. As suggested by Hartshorne, Pogue, and Snedeker (2015), this appears to reflect a tendency to misanalyze fear-type verbs as being frighten-type verbs.

These results are consistent with both the privileged link hypothesis and the type frequency hypothesis. It is unknown how it matches the salience hypothesis, because we do not know whether fear-type or frighten-type event representations are more salient.

Experiment 2: Chase/Flee

Experiment 1 confirmed the prior findings by Hartshorne and colleagues (2015), in which frighten-type verbs were learned earlier than fear-type verbs, despite being lower-frequency. As reviewed above, these admit several difference explanations: frighten-type verbs encode a causal agent as the subject, are more numerous, and arguably encode more salient events than do fear-type verbs.

In order to start disentangling these possibilities, we next considered a different perspective pair: verbs that describe chasing and verbs that describe fleeing. Predictions are summarized in Table 1. As with psych verbs, the class that is the most numerous (chase-type) is also the one where the verb's subject most clearly encodes causality. Thus, if either of these factors explained the early learning of frighten-type verbs, we would expect chase-type verbs to be similarly advantaged. In addition, the chase-type verbs might be early-learned because they arguably encode the more salient event perspective.

One concern about Experiment 1 is that the Truth Value Judgment task may have been more difficult for the youngest children. In particular, describing internal emotional states of the characters requires a fairly involved story (at least, by the standards of stories for 3 year-olds). In contrast, the chase/flee verbs lend themselves naturally to an act-out task – something that was obviously not possible for psych verbs. Thus for Experiment 2, we adopted an act-out task paradigm.

Method

Participants

We recruited 208 children ages 3 through 6 from the Greater Boston Area. All participants were native English speakers. We excluded 41 children for failing to complete more than half of the items or experimentor error. Although we aimed for 40 participants per age (10 per list; see below), we ultimately obtained 43 3 year-olds, 43 4 year-olds, 40 5 year-olds, and 41 6 year-olds.

Materials

Table 4: Verbs used in Experiment 2, with frequency in parts per million and probability that causality is assigned to the sentential subject.

Verb	Type	Frequency	LogFrequency	SubjCause
pursue	chase	0.0	0.0	96
chase	chase	75.9	4.3	92
follow	chase	90.3	4.5	88
flee from	flee	0.3	0.3	57
escape from	flee	4.6	1.7	70
run from	flee	8.5	2.3	61

We selected the 3 highest-frequency chase-type verbs (*chase*, *follow*, *pursue*) and 3 highest-frequency flee-type verbs (*flee*, *escape*, *run*) (Table 4). Frequencies were determined as in Exp. 1, with the caveat that we restricted frequency counts for *run* to specifically the bigram *run from*. Two puppet participants, Giraffe and Tiger, were used for each sentence. We created 6 stories which participants needed to act out with the two puppets. Each story was one sentence long (e.g., *Giraffe chased Tiger*) to minimize demands on working memory and clearly isolate the verb of interest. Giraffe and Tiger were counterbalanced between acting as the subject and the object for each verb. Four lists were created by counter-balancing across lists whether Giraffe or Tiger was the subject of each verb and by creating two item orders, one of which was the reverse of the other.

Procedure

The researchers began by familiarizing the participants with the puppets. Participants were then informed that they would be listening to stories about Giraffe and Tiger, and that they should act out those stories using the puppets. The stories were read aloud, and the researchers recorded whether or not the child correctly demonstrated the meaning of the verb using the puppets. This was operationalized by the child moving

Table 5: Regression results for chase/flee verbs

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.356	0.277	4.897	0.00000
Age	1.002	0.185	5.416	0.00000
VerbType	-1.706	0.502	-3.400	0.001
LogFrequency	1.661	0.316	5.258	0.00000
Age:VerbType	-0.659	0.284	-2.318	0.020
Age:LogFrequency	0.907	0.165	5.502	0.00000
VerbType:LogFrequency	-1.079	0.635	-1.699	0.089
Age:VerbType:LogFrequency	-1.025	0.322	-3.180	0.001

the puppets in the correct direction (i.e. the Giraffe towards the Tiger for "Giraffe chased Tiger") or by reorienting the chaser to be facing the correct direction (i.e. the Giraffe faced the Tiger for "Giraffe chased Tiger"). The on-site record was later checked by another researcher using a video recording of the study.

Results

The dependent measure was accuracy: whether the participant correctly acted out the event (see description of coding procedure above). The mixed effects logistic regression revealed that every main effect and interaction was significant except the interaction of verb type and log frequency, which trended towards significance (Table 5).[12]

Figure 3 plots the main results. While accuracy was roughly similar on both argument structure classes across the age range studied (Fig. 3, top), this belies a large difference in frequency. In particular, although *chase* and *follow* are much higher frequency than *escape from* and *run from*, they are learned no earlier (Fig. 3, middle). As noted above, this not-quite-significant interaction is actually modulated by a significant three-way interaction of age, token frequency, and argument structure class. To aid interpretation of the three-way interaction between verb type, age, and log frequency (Table 5), we plotted marginal effects (Fig. 4). This shows that at a range of frequencies, the inferred rate of learning for flee-type verbs exceeds that of chase-type. Indeed, the inferred rate of learning for *run from* (Log Frequency = 2.3) is actually higher than that of *chase* (Log Frequency = 4.3) or *follow* (Log Frequency = 4.5).

[12]The results of the Bayesian regression were reasonably similar. In particular, the critical three-way interaction of age, log frequency, and verb type was again significant (p = 0.016). The biggest difference was that the main effect of verb type was not significant (p 0.34), nor was the interaction of age and verb type (p 0.13). However, this was less of a difference in the point estimates than just a high degree of uncertainty. The other effects that were significant in the frequentist analyses where either significant in the Bayesian analyses – the two-way interaction of age and log frequency (p <0.00000001) and the main effect of age (p <0.00000001) – or trended towards significance – log frequency (p = 0.074).

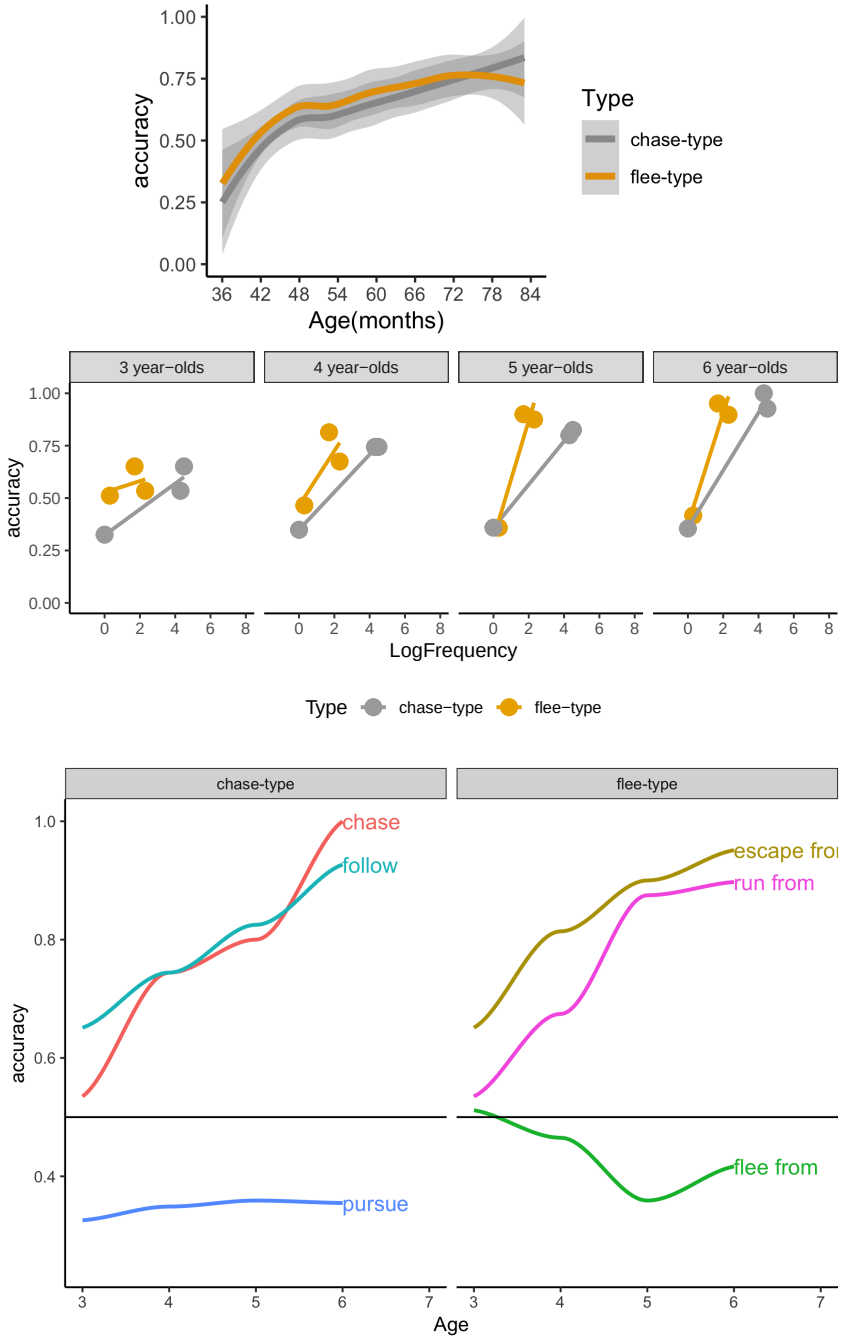


Figure 3. ***Results from Exp. 2. Top: Averaging across verbs within type, with LOESS smoothing over age. Middle: Accuracy for each verb against log frequency, split into four age groups, with linear regressions shown. Bottom: Performance on each verb across ages, aggregated into four age groups, with LOESS smoothing.***

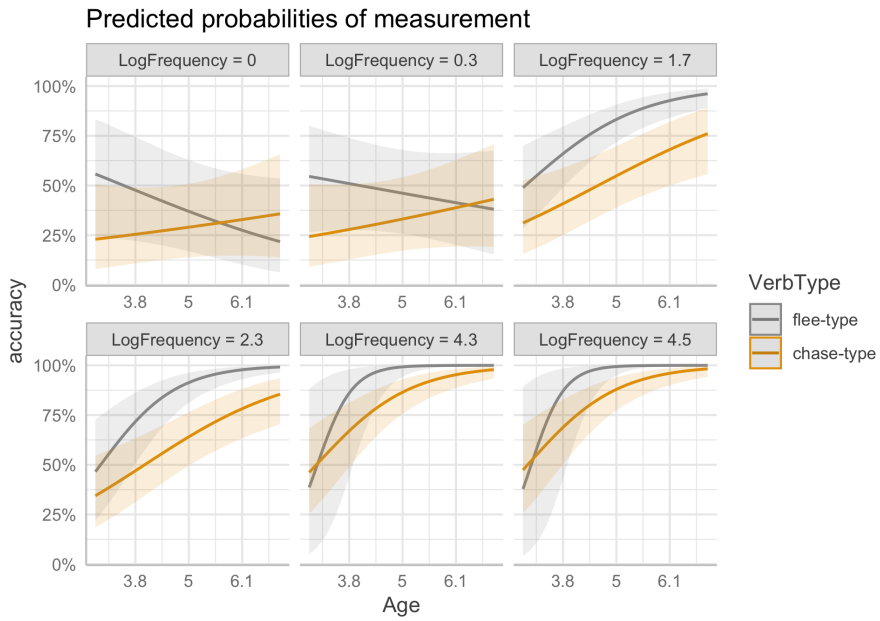


Figure 4. ***Marginal effects for the interaction of verb type, (z-scored) age, and frequency in chase/flee verbs (Exp. 2). Note that the curves plotted are inferred by the model; they do not represent direct observations. For instance, there was no flee-type verb with a log frequency of 0.***

As explained in "Overview of Analyses," we followed up this analysis by considering each age group separately. For 3 year-olds, neither the main effects nor their interaction were significant (*p*s $\geq$ 0.11). For 4 year-olds, there were significant main effects of log frequency (B = 1.30, SE = 0.43, Wald's z = 3, *p*=0.0025) and verb type (B = -1.40, SE = 0.69, Wald's z = -2, *p*=0.045), though their interaction was not significant (B = -0.50, SE = 0.86, Wald's z = -0.59, *p*=0.56). At age 5, there remained significant main effects of log frequency (B = 2.90, SE = 0.64, Wald's z = 4.50, *p*=0.0000075) and verb type (B = -2.70, SE = 1.10, Wald's z = -2.40, *p*=0.015), and the interaction approached significance (B = -2.30, SE = 1.30, Wald's z = -1.80, *p*=0.076). By the age of 6, all effects were significant: log frequency (B = 15, SE = 3.50, Wald's z = 4.20, *p*=0.000029), verb type (B = -13, SE = 4.20, Wald's z = -3, *p*=0.0029), and their interaction (B = -24, SE = 5.80, Wald's z = -4.10, *p*=0.000042).

Discussion

The results of Exp. 2 indicate an early advantage of flee-type verbs relative to chase-type verbs, holding frequency constant, albeit perhaps not as pronounced as that for frighten-type verbs relative to fear-type verbs. All three of our main hypotheses predicted an advantage for chase-type verbs, though in the case of the salience hypothesis, the evidence driving this prediction is weak. In any case, the results were exactly the opposite.

Interestingly, one chase-type verb and one flee-type verb each exhibited stubbornly sub-chance levels: Children responded as if *pursue* was flee-type and *flee from* was chase-type – something that had not resolved even by the age of 6. This is puzzling, and we do not have much to say about it at the moment other than that it is not obviously explicable under any of the hypotheses being considered.

It should be noted that the observed differences between chase-type and flee-type verbs are due primarily to four verbs (the differences in learning outcomes for *pursue* and *flee from* are fairly similar). Thus, different results for a single verb would have substantially changed the statistical outcomes. Unfortunately, because flee-type verbs are so rare type-wise, not much can be done about this other than to consider other perspective pairs as well.

Experiment 3

The results of Experiment 2 converged with those of Experiment 1 in that token frequency was a poor predictor of learning. Otherwise, the results contrast: unlike Experiment 1, the results of Experiment 2 were inconsistent with the suggestion that acquisition is heavily influenced by class token frequency or causality. The results moreover suggested that salience plays little role.

Nonetheless, Experiment 2 considered a very small number of verbs. Thus, in Experiment 3, we turn to another case study: give/get verbs. In addition to being more numerous, give/get verbs offer the advantage of disentangling the predictions of the causal semantics and token frequency hypotheses on the one hand from those of the salience hypothesis on the other (Table 1).

While give and get events can be acted out, it is a bit more complicated to manage with small hands, as there are three entities to keep track of. Moreover, we wished to avoid providing clues to the direction of motion in the form of prepositions (giving *to* vs. getting *from*). Thus, we used a modified video-description task: children watch an event involving a boy and a girl exchanging items and were queried as to what the one of the characters got/gave/etc.

Method

Participants

We recruited 452 native English-speaking children ages 36 to 83 months old from the Greater Boston Area. We excluded 28 children due to experimentor error or failure to complete more than half the trials. While we had intended to recruit 9 participants per age group per list (see below), we fell 8 participants short due to recruitment restrictions. Thus, we finished with 424: 107 3 year-olds, 105 4 year-olds, 109 5 year-olds, and 104 6

year-olds.

Materials

The four give-type verbs and four get-type verbs were selected from the Verbnet Unified Verb Index. These were the four highest-frequency verbs that we were able to identify for each type (Table 6). The two groups did not differ significantly in terms of frequency ($\Delta M = 0.17$, 95% CI $[-4.32, 4.67]$, $t(4.34) = 0.10$, $p = .921$).

Table 6: Verbs used in Experiment 3, with frequency in parts per million and probability that causality is assigned to the sentential subject.

Verb	Type	Frequency	LogFrequency	SubjCause
receive	get	4.6	1.7	4
grab	get	65.4	4.2	96
buy	get	423.4	6.1	58
get	get	6944.0	8.8	26
sell	give	62.1	4.1	73
send	give	75.2	4.3	99
pass	give	85.7	4.5	94
give	give	1318.2	7.2	96

We constructed videos depicting one male and one female experimenter exchanging objects with one another. Fig. 5 shows an example: the man gives an apple to the woman, who then reciprocates by giving a hammer to the man. The participant would then be asked either *what did the boy give?* or *what did the girl give?* The fact that we could query either character allowed us to counter-balance SOURCEs and GOALs within the same stimuli. We counter-balanced another way as well: we made pairs of videos that differed in which character moved first, which should wash out any bias towards the first- (or last-) mover. Similarly, the videos were designed to allow querying one give-type verb and one get-type verb: the same video was used for either *give* or *get*, another for *send* or *receive*, and yet another for *buy* and *sell*. An exception to this design was forced by *grab*, which was paired with *pass*. For obvious reasons, the same video could not be used for both. Moreover, we found videos with reciprocal grabbing to be confusing. Thus, rather than have the man and woman both grab from each other within the same video, we created two videos – one with grabbing by the man and one with grabbing by the woman – and placed them one above the other. Instead of counter-balancing which character acted first, we counter-balanced which video was on top. Because *pass* was matched with *grab* for purposes of counter-balancing, we constructed the *pass* stimuli in a similar way.

Four lists were constructed by counterbalancing the order in which the verbs were

Figure 5. ***Stills from one of the videos for*** **give**, ***used in Exp. 3.***

queried and, for a given verb, which character was queried and which character moved first in the video. To achieve this, we used a single fixed order of the videos (which a caveat describes below), counter-balancing which character is asked about and which verb in each verb pair was queried. The caveat is that this was obviously not possible for *pass* and *grab*. Instead, for these we swapped the order of the videos (so on two lists, *pass* was presented before *grab*, and the reverse was true for the other two lists) and the vertical positioning of the videos (see above).

Finally, these four lists were triplicated by making two more sets of videos, each with a different pair of actors and different sets of objects, for a total of 12 lists.

Procedure

Videos were presented on an iPad using Keynote. Prior to watching each video, participants viewed the opening frame and were asked to point to the target objects, with corrections provided as necessary. This helped ensure they could identify the objects well enough to answer the subsequent questions. The video was then played, and the participant was asked the question involving the give/get verb for that video.

Results

The dependent measure was accuracy: whether the participant named the correct item. The three-way interaction of age, log frequency, and verb type was significant, as was the interaction of age and verb type and the main effects of frequency and age (Table 7).[13] The data plots provide context. Averaging within verb-type, there is a clear early advantage for get-type verbs, which disappears by age 6 in part because performance improves with age for give-type but not get-type verbs (Fig. 6 top). Plots of individual verbs (Fig. 6 bottom) show steady improvement with age for all give-type verbs, with

[13]The three-way interaction that was significant in the frequentist analyses only trended towards significance in the Bayesian regression (p=0.08). The interaction of age and verb type was again fully significant (p=<0.00000001), as was the main effect of age (p=<0.00000001). The only effect that was significant in the frequentist analyses that was was clearly not significant in the Bayesian analyses was the main effect of log frequency (p=0.19), though as usual this reflected high uncertainty rather than certainty that there is no effect.

Table 7: Regression results for give/get verbs

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.605	0.310	5.171	0.00000
Age	0.615	0.064	9.548	0
VerbType	0.100	0.623	0.161	0.872
LogFrequency	0.985	0.392	2.512	0.012
Age:VerbType	0.728	0.141	5.157	0.00000
Age:LogFrequency	0.044	0.076	0.575	0.565
VerbType:LogFrequency	0.292	0.784	0.372	0.710
Age:VerbType:LogFrequency	-0.344	0.153	-2.253	0.024

much less improvement for three of the get-type verbs and a substantial decline in performance for *receive*.

Fig. 6 provides some context for the three-way interaction between age, log frequency, and verb type: There is initially an advantage for get-type verbs – particularly low-frequency ones – but that dissipates by ages 4 and 5 and may even start to reverse by age 6. These qualitative observations were largely confirmed by quantitative analysis. For 3 year-olds, the main effect of log frequency was significant (B = 0.92, SE = 0.30, Wald's z = 3.10, p=0.002) and the main effect of verb type trended towards significance (B = -0.86, SE = 0.49, Wald's z = -1.80, p=0.077), though their interaction was not significant (B = 0.91, SE = 0.60, Wald's z = 1.50, p=0.13). At ages 4 and 5, the only effects to even approach significance were the significant effects of log frequency (4 year-olds: B = 1.10, SE = 0.47, Wald's z = 2.40, p=0.017; 5 year-olds: B = 1.10, SE = 0.50, Wald's z = 2.20, p=0.028). At the age of 6, no effects or interactions were significant (ps $\geq$ 0.14, SE = 0.95, Wald's z = 0.42, p=0.67).

Discussion

As with psych verbs and chase/flee verbs, the give/get verbs showed a clear effect of verb type. However, it was in many ways the opposite of what was observed for the first two perspective pairs: rather than an advantage for one verb-type emerging between the ages of 3 and 6, we observed an early advantage for get-type verbs that dissipated.

It may not be necessary to make much of this difference. Three year-olds showed much higher accuracy on give-type and especially get-type verbs than they did on either type of psych verbs or chase/flee verbs – something that is consistent with their relatively early emergence in spontaneous speech (Bowerman, 1990) and extremely high token frequency. Get-type verbs must have diverged from give-type verbs at an earlier age, and had we tested two year-olds, our results may have looked (more) qualitatively similar to

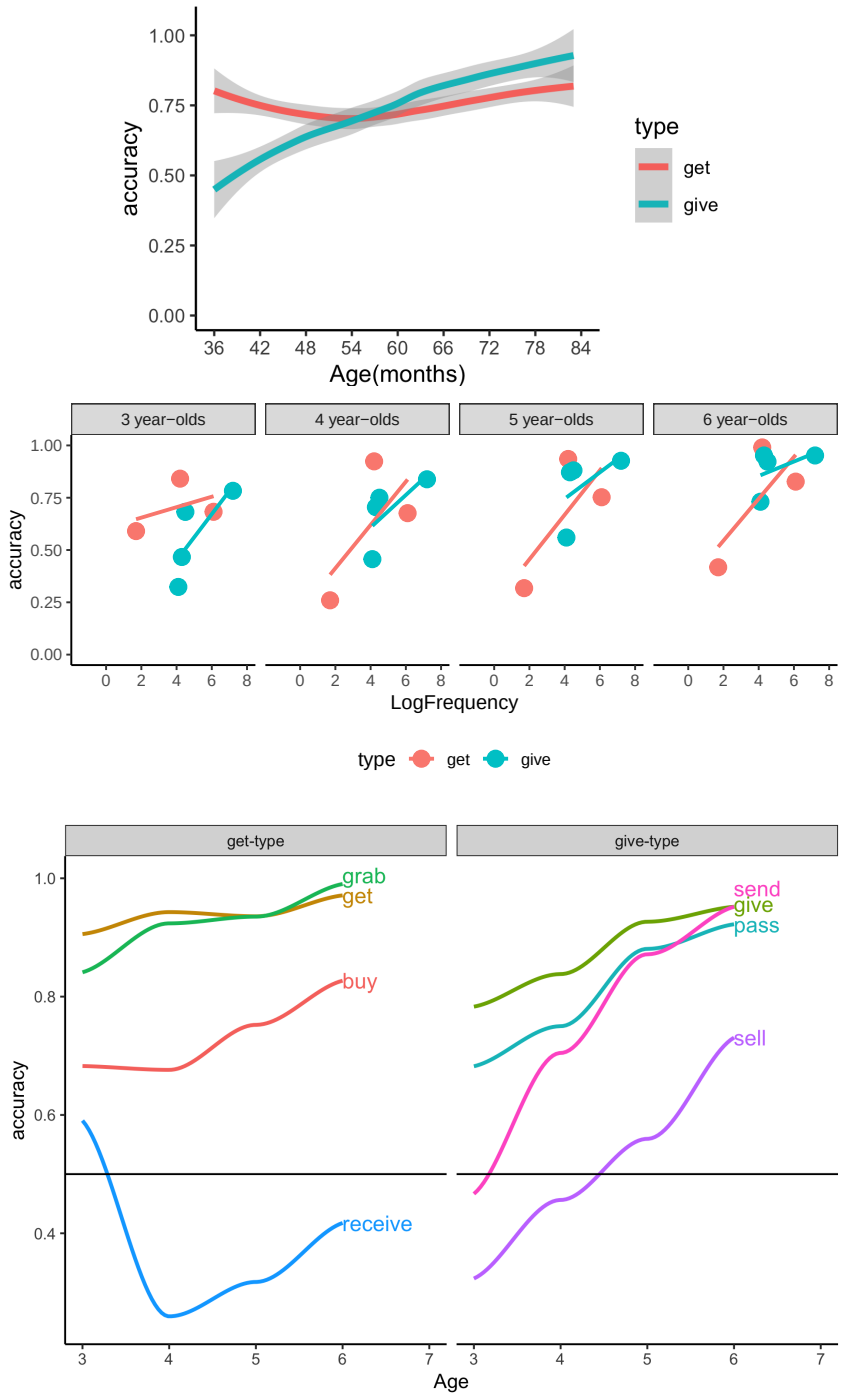


Figure 6. ***Results from Exp. 3. Top: Averaging across verbs within type, with LOESS smoothing over age. Middle: Accuracy for each verb against log frequency, split into four age groups, with linear regressions shown. Bottom: Performance on each verb across ages, aggregated into four age groups, with LOESS smoothing.***

those of the other two perspective pairs, simply shifted earlier in development.

Regardless, the findings again contradict the type-frequency hypothesis and the privileged link hypotheses, both of which predicted an advantage for give-type verbs. The privileged link hypothesis fails even if we take into account variation in the causal semantics of get-type verbs, the only class to show much variation (see Figure 7, right).

The results are, however, broadly consistent with the salience hypothesis. However, there are the puzzling U-shaped results for *receive*. Given the very large number of participants, statistical fluke seems unlikely. One possibility is that this reflects a misanalysis of the meaning of *receive*, much like what we observed for some fear-type verbs. Indeed, of all verbs in the dataset, participants in our rating study were most likely to rate *receive*'s object as causally responsible (89). Under the privileged link hypothesis, this could result in learners mistakenly treating *receive* as synonymous with *give*. Unfortunately, the rest of our results are not kind to the privileged link hypothesis, leaving this finding as something of a mystery.

General discussion

We investigated the development of high-frequency verbs in three "perspective pair" classes: emotion (psych) verbs, chase/flee verbs, and give/get verbs. In each case, argument structure type was predictive of learning, above and beyond token frequency. Taking token frequency into account, we found unexpectedly early acquisition of frighten-type verbs relative to fear-type verbs, of flee-type verbs relative to chase-type verbs, and of get-type verbs relative to give-type verbs (though by four years of age, this last difference had dissipated). Indeed, collapsing across experiments, while there is an effect of token frequency, it is clearly modulated by large effects of verb class (Figure 7, top).

This is not to say that token frequency and argument structure class are the only predictors of order of acquisition. There is additional variability not captured by these constructions (at least, to the degree we can accurately measure either; see 'Methodological Limitations', below). However, the effect was similar in size and reliability to that of token frequency, making it unusually potent by the standards of psychology.

Similarly, we cannot be sure that the verbs' argument structure causally affected pace of acquisition or was merely correlated with something that did. The obvious ways to either explain the effect of argument structure or explain it away did not pan out. The order of verb acquisition appears to be even less affected by degree to which the verb describes an event caused by the subject. Indeed, the class of verbs with greater subject causality was learned later in two of three cases (again, excluding only the psych verbs). Collapsing across all three experiments, there is little evidence of a relationship

between these variables (Figure 7, middle).

Neither is the order of acquisition much predicted by the type frequency of its argument structure class. This was clear in the analyses above, where the class with greater type frequency was actually learned less well in two out of three cases (all except psych verbs). The lack of a relationship between type frequency and acquisition is made even more clear in the summary figure, which collapses across experiments (Figure 7, bottom).

The salience hypothesis is least wrong, primarily by virtue of not making many clear predictions, at least at present. It makes no predictions about fear/frighten verbs, since we currently do not know which perspective is more salient. It correctly predicts the earlier learning of get-type verbs – at least, if one adopts our interpretation of prior work on goal salience (see above). As we reviewed in the Introduction, there is currently some suspicion that the "chase" perspective is more salient than the "flee" perspective, which would be inconsistent with our finding of early learning of flee-type verbs. However, that suspicion is based on sparse data with inconsistent results. So one could reasonably argue that we have no predictions about chase/flee verbs, either.

Interestingly, all three experiments showed some evidence of U-shaped learning. U-shaped learning could be interpreted as a verb being mislearned as belonging to the wrong class. Thus, if U-shaped learning was specific to the classes that were learned later, it would be evidence of the earlier-learned classes being in some way more salient or easier to apply. However, the pattern was unclear. While four fear-type verbs showed U-shaped or below-chance learning (*trust, dislike, enjoy, fear*) – thus replicating and extending earlier observations by Hartshorne, Pogue, and Snedeker (2015) – so did two frighten-type verbs (*anger, amaze*). Among chase/flee verbs, one of each type showed below-chance learning (*pursue, flee*). Among give/get verbs, there was only one example of U-shaped learning, but it was a member of the get-type class (*receive*), which was overall acquired earlier.[14]

Methodological limitations

Before further discussing the theoretical implications, we consider several limitations in the data and the strength of the evidence.

Most broadly, we only investigated the acquisition of 30 verbs across three perspective pair classes in a single language. This is certainly a substantial improvement on prior work, which focused on narrow age ranges and dealt primarily with fear/frighten verbs in English (Bowerman, 1990; Braine et al., 1993; Hartshorne et al., 2015; Messenger et

[14]We also considered the possibility that these were the verbs for which the object was relatively causal. Indeed, the verb with strongest causality ratings for the object was *receive* (89). However, *pursue* was among those with the least causal ratings for the object (0). On the whole, there was no clear relationship.

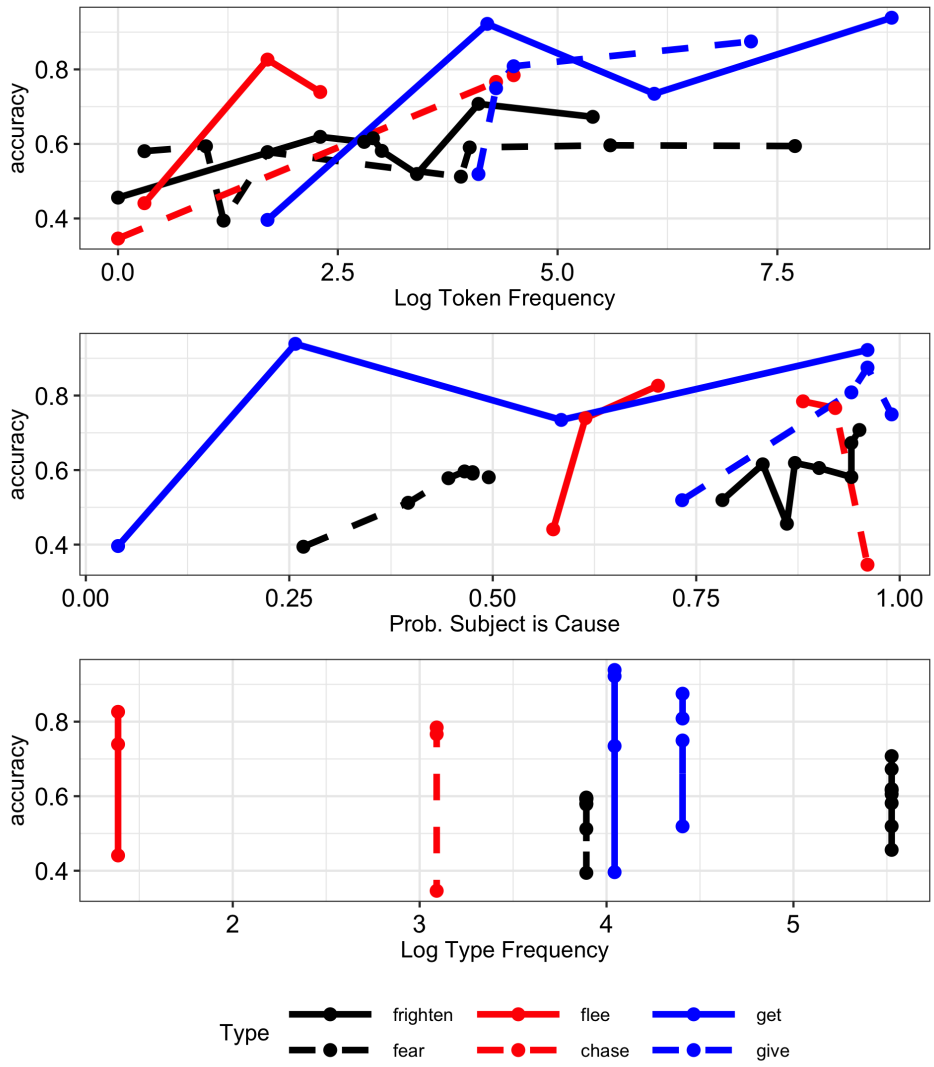


Figure 7. ***Quantitative results, summarizing across experiments. Top: Likelihood of correctly understanding a verb's argument structure is significantly if only moderately predicted by token frequency (b = 0.10, 95% CI = [0.05, 0.14], t(28) = 4.26, p < .001, R^2 = 0.39). Center: The degree to which the subject of the verb causes the event predicted accuracy even less well, an effect that approached but did not reach significance (b = 0.05, 95% CI = [-0.01, 0.11], t(28) = 1.83, p = .078; R^2 = 0.11). Bottom: There was no evidence of a relationship between accuracy and class type frequency (b = -0.02, 95% CI = [-0.08, 0.04], t(28) = -0.61, p = .545; R^2 = 0.01). Note: To facilitate comparison, the above regression coefficients are standardized.***

al., 2012; Tinker et al., 1988), with the exception of one study of the emergence in the spontaneous speech of two children of four give/get verbs (Bowerman, 1990). A few additional studies looked at relative preference for one perspective pair or the other in elicited naming, with mixed results (Fisher et al., 1994; Gleitman et al., 2007; e.g., Lakusta & Landau, 2005). While our studies included most of the verbs in the three perspective pair classes that were sufficiently high frequency in child-directed speech to be plausibly known by young children, it is quite possible the results would be different if other verbs had been available. Moreover, the results of these three perspective pairs may be unrepresentative even within English, and may not generalize to other languages – particularly languages that organize argument structure differently, such as ergative or agglutinative languages. It would certainly not be the first time such generalization failed (Evans & Levinson, 2009; Yarkoni, 2019).

Related to that point, while the sample sizes are not small by the standards of language development research, they are nonetheless not large by the standards of statistical analysis, particularly when investigating three-way interactions (Hartshorne & Schachner, 2012; Vankov, Bowers, & Munafò, 2014). Some comfort is given by the fact that in all three experiments, the key interactions that were significant in our frequentist analyses were also significant or trended towards significance in the more conservative Bayesian analyses. (The lone exception was the interaction of argument structure class and log frequency in Exp. 1.) However, interpretation of the p-values we report must take into account the fact that the small number of verbs precluded fitting maximal random effects, which may or may not be anti-conservative (Barr et al., 2013; Bates et al., 2015; Matuschek et al., 2017).

Similarly, while the estimates of frequency in child-directed speech are based on the largest dataset available (all part-of-speech-tagged corpora involving native English-speaking, North American children in CHILDES-db in the focused age range) the resulting dataset is not actually very large – around 3 million tokens, which is far less than what one child hears in a single year (Hart & Risley, 1995) and is likely to be biased by the sampling strategies used by the researchers. Moreover, many of the part-of-speech tags are automatically generated and of variable quality. Additionally, with the exception of *run from*, we counted all uses of the verbs, rather than uses in the critical syntactic frames. This was driven by a practical consideration (automatically extracting syntactic frames is difficult, particularly for spoken corpora), but determining whether it is reasonable theoretically will require more exact theories of language acquisition than we currently have.

The three experiments use three different methods: Truth Value Judgment, act-out, and question-answering. These differences were driven by the semantics of the verbs, which rendered different methods more or less natural. Consider, for instance, how one would run an act-out task for fear-type verbs, which describe a habitual disposition, not any particular action. Nonetheless, this methodological variation limits direct cross-study

comparison.

Relatedly, the chase/flee experiment has a potential confound in that flee-type verbs have prepositions but chase-type verbs do not. In principle, children might have responded by ignoring the verb and focusing on the preposition. This can, of course, cut both ways: the particular sentence frame used by chase-type verbs is relatively rare compared to the transitive, which may make it more difficult for children to comprehend. This consideration just further highlights the need for data on more different kinds of perspective pair verbs.

Our test of the 'privileged link' hypothesis is contingent on our operationalization of causality. Specifically, we asked participants whether each event participant "made" the event happen. There are other ways to operationalize causality that might lead to different results. For instance, Hartshorne et al. (2016) embedded judgments of causality in a legal context, which they found resulted in particularly sharp judgments. On the other end, some authors have argued that causality is too narrow a category, preferring broader notions such as "acting on" (MacWhinney, 1977). (An anonymous reviewer has suggested we consider "actively doing something to.") More generally, the definition of "cause" is contentious even outside the linguistic domain (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021). It is unclear whether a different operationalization or causality or the selection of a different cause-related construct such as "act on" would have resulted in substantially different findings; presumably, all these constructs are reasonably correlated. Nonetheless, it will be impossible to completely rule out this worry at least until we have a full understanding of semantic representations.

Finally, as noted in the Introduction, the present studies test whether participants understood who did to whom. They do not directly address *what* was done. It is possible that participants succeeded on these tasks without understanding the differences between *love* and *hate* or between *grab* and *get*.

Theoretical implications

With the above caveats, the clearest finding is that it is not clear what drives successful verb learning. Learning was not well-predicted by token frequency, causality, or type frequency of the verb's argument class. The best that can be said is that the salience hypothesis was not entirely disconfirmed.

Thus, one obvious direction for future research is to obtain a clearer understanding of which perspective is more salient for different perspective pairs, in order to better test the hypothesis. This might be aided by developing a more nuanced, precise version of the hypothesis. Salience is a phenomenon that requires its own explanation. Perhaps a perspective is more salient because it is more simply represented (chasing involves a simple goal of being where the target is, whereas the goal of fleeing involves a negation:

not being where the pursuer is), because it is more temporally concentrated (frightening happens at a distinct time and place, whereas fearing is an ongoing state of affairs), or because of a recency effect (*get* highlights the end of an event, whereas *give* highlights its beginning; cf. Regier and Zheng (2007)). A more precise account would allow us to make predictions without necessarily having to obtain direct evidence of salience (i.e., figuring out how to measure the salience of hard-to-depict event perspectives like FEAR). For instance, researchers working on the psychophysics of action perception found it fairly straightforward to make artificial agents that chase (a simple heat-seeking policy works quite well), whereas designing artificial agents that could flee effectively required a much more complex policy (Tang et al., 2021). While this actually makes the wrong prediction in the present study (chase-type verbs were learned later, not earlier), a well-defined, quantitative simplicity-based theory of salience is potentially within reach.

Another direction would be to better characterize the quality of learning opportunities. Our findings are based on acquiring verbs earlier or later than might be expected based on input frequency. This implicitly assumes that all encounters with a verb are equally informative, which is not the case. As we reviewed above, Medina and colleagues (2011) argue that word learning is primarily driven by the rare highly informative encounter. While they present a method for identifying these 'eureka' moments, it requires time-intensive hand annotation. Currently-available annotated corpora are vanishingly small; developing one large enough to test whether frighten-type, flee-type, and get-type verbs have more than their fair share of eureka moments will either require an enormous amount of work or some mechanism for automating the annotation (e.g., through machine learning). Note, moreover, Hartshorne, Pogue, and Snedeker (2015) raise some reasons for being skeptical about this explanation, at least with respect to fear/frighten verbs. They suspect that because frighten-type events are ephemeral, speakers are unlikely to remark upon them as they happen, whereas because fear-type states are ongoing, the reverse may be true for them. As a result, it may be easier for children to connect a fear-type utterance with its co-temperal referent than a frighten-type utterance with its non-co-temporal referent.

Encounters with verbs can be more or less informative in other ways as well. Hartshorne, Pogue, and Snedeker (2015) reported that high-frequency fear-type verbs such as *like*, *love*, and *hate* occur primarily with one of a small number of subjects – mostly *I* and *you*. They note that this might induce children to treat these high-frequency bigrams as set phrases. As a result, most uses of these verbs would fail to provide much information about their argument structure. In that case, the frequency analyses above should be redone excluding those high-frequency bigrams. This is not trivial – for one thing, it requires a principled way of determining which bigrams are of sufficiently high frequency – and we leave it to future work.

More generally, verbs differ in many ways beyond argument structure class, token fre-

quency, type frequency, causal structure, and salience. As already noted, an anonymous reviewer suggests that in our data, learning seems to be somewhat earlier for verbs involving events where one or more participants are particularly active and intentional. One can certainly imagine this makes the events more cognitively salient or simply easier to spot in the world. Other cognitive biases, such as a bias towards positive (or negative) events could also play a role.[15] There are currently a vast range of possibilities to explore, given that the ones most grounded in the literature are less explanatory than anticipated.

Conclusion

Hartshorne, Pogue, and Snedeker (2015) reported a puzzling finding: relatively old children failed to understand extremely high-frequency fear-type verbs long after they had acquired a number of lower-frequency frighten-type verbs. They proposed a number of possible explanations based on current understanding of verb-learning. Of the ones that are currently testable, none are clearly consistent with the present results. More broadly, the current study revealed that this is not a funny fact about fear/frighten verbs, but may in fact be a common phenomenon – one that has gone largely undetected and remains essentially without explanation. This conclusion is based on only three case studies in a single language: the empirical picture may be even more complicated than it appears so far. All we can say at the moment is that something is going on, and we do not understand it. This should concern us, because if we are missing a large part of the empirical description of language acquisition, our theorizing may be entirely misdirected. At the very least, it is incomplete. There is more in heaven and earth than is dreamt of in our philosophy, so some new philosophy is needed.

[15]We thank an anonymous reviewer for this specific suggestion.

References

Ambridge, B., Barak, L., Wonnacott, E., Bannard, C., Sala, G., Zwaan, R., & Ferreira, F. (2018). Effects of both preemption and entrenchment in the retreat from verb overgeneralization errors: Four reanalyses, an extended replication, and a meta-analytic synthesis. *Collabra: Psychology*, *4*(1).

Ambridge, B., Maitreyee, R., Tatsumi, T., Doherty, L., Zicherman, S., Pedro, P. M., et al.others. (2020). The crosslinguistic acquisition of sentence structure: Computational modeling and grammaticality judgments from adult and child speakers of english, japanese, hindi, hebrew and k'iche'. *Cognition*, *202*, 104310.

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

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*(3), 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv Preprint arXiv:1506.04967*.

Belletti, A., & Rizzi, L. (1988). Psych-verbs and θ-theory. *Natural Language & Linguistic Theory*, 291–352.

Bowerman, M. (1990). Mapping thematic roles onto syntactic functions: Are children helped by innate linking rules? *Linguistics*, *28*(6), 1253–1289.

Braine, M. D. (1992). What sort of innate structure is needed to "bootstrap" into syntax? *Cognition*, *45*(1), 77–100.

Braine, M. D., Brooks, P. J., Cowan, N., Samuels, M. C., & Tamis-LeMonda, C. (1993). The development of categories at the semantics/syntax interface. *Cognitive Development*, *8*(4), 465–494.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychome-*

trika, *78*(4), 685–709. https://doi.org/10.1007/s11336-013-9328-2

Dowty, D. R. (1989). On the semantic content of the notion of "thematic role." In *Properties, types and meaning* (pp. 69–129). Springer.

Dowty, D. R. (1991). Thematic proto-roles and argument selection. *Language*, *67*(3), 547–619.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–448.

Fisher, C., Gertner, Y., Scott, R. M., & Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(2), 143–149.

Fisher, C., Hall, D. G., Rakowitz, S., & Gleitman, L. R. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua*, *92*, 333–375.

Freeman, N. H., Sinha, C. G., & Stedmon, J. A. (1981). The allative bias in three-year-olds is almost proof against task naturalness. *Journal of Child Language*, *8*(2), 283–296.

Fujita, H. I. I. (2000). A cognitive approach to errors in case marking in japanese agrammatism. *Constructions in Cognitive Linguistics: Selected Papers from the Fifth International Cognitive Linguistics Conference, Amsterdam, 1997*, *178*, 123. John Benjamins Publishing.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading Technical Report; No. 257*.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936.

Gleitman, L. R. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, *57*(4), 544–569.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*.

Oxford University Press on Demand.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language, 35*(3), 515–531.

Grimshaw, J. (1990). *Argument structure.* the MIT Press.

Hansen, P. (2017). What makes a word easy to acquire? The effects of word class, frequency, imageability and phonological neighbourhood density on lexical development. *First Language, 37*(2), 205–225.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children.* Paul H Brookes Publishing.

Hartshorne, J. K., Bonial, C., & Palmer, M. (2014). The VerbCorner project: Findings from phase 1 of crowd-sourcing a semantic decomposition of verbs. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 397–402.

Hartshorne, J. K., O'Donnell, T. J., Sudo, Y., Uruwashi, M., Lee, M., & Snedeker, J. (2016). Psych verbs, the linking problem, and the acquisition of language. *Cognition, 157*, 268–288.

Hartshorne, J. K., Pogue, A., & Snedeker, J. (2015). Love is hard to understand: The relationship between transitivity and caused events in the acquisition of emotion verbs. *Journal of Child Language, 42*(3), 467.

Hartshorne, J. K., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience, 6*, 8.

Hlavac, M. (2018). *Stargazer: Well-formatted regression and summary statistics tables.* Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). Retrieved from https://CRAN.R-project.org/package=stargazer

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2008). A large-scale classification of english verbs. *Language Resources and Evaluation, 42*(1), 21–40.

Lakusta, L., & Landau, B. (2005). Starting at the end: The importance of goals in spatial language. *Cognition, 96*(1), 1–33.

Lakusta, L., & Landau, B. (2012). Language and memory for motion events: Origins of the asymmetry between source and goal paths. *Cognitive Science, 36*(3), 517–544.

Landau, B., & Gleitman, L. R. (2015). 10 height matters. *Structures in the Mind: Essays on Language, Music, and Cognition in Honor of Ray Jackendoff*, 187.

Levin, B., & Hovav, M. R. (2005). *Argument realization*.

Lidz, J., Gleitman, H., & Gleitman, L. R. (2003). Understanding how input matters: Verb learning and the footprint of universal grammar. *Cognition*, *87*(3), 151–178.

Lüdecke, D. (2018). Ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), 772. https://doi.org/10.21105/joss.00772

MacWhinney, B. (1977). Starting points. *Language*, 152–168.

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

Marantz, A. (1982). On the acquisition of grammatical relations in linguistik als kognitive wissenschaft. *Linguistische Berichte Braunschweig*, (80), 32–69.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type i error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McDonough, C., Song, L., Hirsh-Pasek, K., Golinkoff, R. M., & Lannon, R. (2011). An image is worth a thousand words: Why nouns tend to dominate verbs in early word learning. *Developmental Science*, *14*(2), 181–189.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019.

Messenger, K., Branigan, H. P., McLean, J. F., & Sorace, A. (2012). Is young children's passive syntax semantically constrained? Evidence from syntactic priming. *Journal of Memory and Language*, *66*(4), 568–587.

Papafragou, A. (2010). Source-goal asymmetries in motion representation: Implications for language production and comprehension. *Cognitive Science*, *34*(6), 1064–1092.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, Mass.: Harvard University Press.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure.*

Quine, W. (1964). *Word and object.*

Regier, T., & Zheng, M. (2007). Attention to endpoints: A cross-linguistic constraint on spatial meaning. *Cognitive Science*, *31*(4), 705–719.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). Childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, *51*(4), 1928–1941.

Strickland, B., Fisher, M., Keil, F., & Knobe, J. (2014). Syntax and intentionality: An automatic link between language and theory-of-mind. *Cognition*, *133*(1), 249–261.

Tang, N., Gong, S., Liao, Z., Xu, H., Zhou, J., Shen, M., & Gao, T. (2021). Jointly perceiving physics and mind: Motion, force and intention. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*.

Tinker, E., Beckwith, R., & Dougherty, R. (1988). Markedness and the acquisition of emotion verbs.

Vankov, I., Bowers, J., & Munafò, M. R. (2014). Article commentary: On the persistence of low power in psychological science. *Quarterly Journal of Experimental Psychology*, *67*(5), 1037–1040.

Wellman, H. M. (1992). *The child's theory of mind.* The MIT Press.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.org/knitr/

Yarkoni, T. (2019). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37.

Data, Code and Materials Availability Statement

Data, code, and materials are available at https://osf.io/k5xud/.

Ethics statement

Ethics approval was obtained from the ethics committee of the University of Nowhere. Parents of child participants gave gave informed written consent before the child took part in the study. Adult participants recruited and tested online read a consent statement before participating.

Authorship and Contributorship Statement

JKH conceived of the study. JKH and LS designed the experiments. JKH, YH, and LS performed analyses. JKH and YH wrote the manuscript.

Acknowledgements

The authors thank Jesse Snedeker, Tim O'Donnell, Laura Lakusta, Vera Kempe, and two anonymous reviewers for helpful feedback and suggestions, and Miguel Mejia, Zach Barker, Kayley Okst, Caitlin Garcia, Nicola Roux, Marissa Russell, Camille Phaneuf, Ernesto Gutierrez, Juliani Vidal, Hayley Greenough, Casey Nicastri, Rudmila Rashid, Taylor Martinz, Kate Roberts, Madeleine McCanne, Sarah Al-Mayahi, Alice Lim, Everett Kim, Lily Feinberg, Zeen Naeem and the staff at the Acton Discovery Museum and Boston Children's Museum for assistance with data-collection. Funding was provided by NSF 1551834, awarded to JKH.

License

Ad-hoc pragmatic implicatures among Shipibo-Konibo children in the Peruvian Amazon

Martin Fortier
PSL Research University, France

Danielle Kellier
University of Pennsylvania, USA

María Fernández Flecha
Pontificia Universidad Católica, Peru

Michael C. Frank
Stanford University, USA

Abstract: Pragmatic reasoning – the ability to infer the intended meaning of an utterance in context – is one of the core aspects of language comprehension. Children's ability to reason pragmatically increases across childhood in U.S. and European communities. In these communities, ad-hoc (contextual) implicatures tend to emerge around age four, but this pattern has not been studied across a broader range of contexts. We conducted a study of the development of ad-hoc implicatures in Shipibo-Konibo communities in the Peruvian Amazon. While 8–11-year-olds successfully made ad-hoc implicatures, younger children did not, despite successfully understanding control trials. These findings suggest that ad-hoc implicatures are available interpretations but that their development may be more protracted.

Keywords: Shipibo-Konibo; pragmatic development; implicature.

Corresponding author(s): Michael C. Frank, Department of Psychology, Stanford University, 450 Jane Stanford Way, Stanford, CA 94305, USA. Email: mcfrank@stanford.edu.

ORCID ID(s): https://orcid.org/0000-0002-7551-4378

Citation: Fortier, M., Kellier, D., Fernández-Flecha, M. & Frank, M. C. (2023). Ad-hoc pragmatic implicatures among Shipibo-Konibo children in the Peruvian Amazon. *Language Development Research, 3*(1), 105—120. https://doi.org/10.34842/2023.76

Introduction

One of the most salient – and fascinating – aspects of human language is the ability of speakers to express a complex and subtle range of meanings that go beyond the literal semantics of their utterances. To take a classic example, a letter of recommendation that contains praise for penmanship and punctuality can still be damning based on its omission of certain important information (Grice, 1975). This kind of pragmatic reasoning – reasoning about a speaker's intended meaning in a particular context – allows the flexible, social use of language to accomplish a wide variety of different goals (Grice, 1975; Clark, 1996; Goodman & Frank, 2016).

In the current study, we use ad-hoc implicatures – a specific pragmatic phenomenon – as a case study of the broader process of contextual reasoning about language. For example, in the letter of recommendation case, the pragmatic implicature is that the candidate is *not* intelligent or hard-working. Because such implicatures can be constructed even in very simple contexts, they can be a useful tool for studying developmental change in pragmatic inferencing ability (e.g., Papafragou & Tantalou, 2004; Stiller et al., 2015; Horowitz et al., 2018). But to date this research has only been conducted in WEIRD – western, educated, industrialized, rich, democratic – contexts (Henrich et al., 2010). We extend this work by examining developmental change in ad-hoc implicatures in a non-WEIRD culture, the Shipibo-Konibo (SK) people of the Peruvian Amazon.[1]

In the remainder of this introduction, we introduce questions about the cross-cultural universality of pragmatic principles, the current state of the developmental evidence, and the specifics of our investigation.

Gricean Pragmatics Across Cultures

Grice's (1975, 1981) theory of pragmatics and implicatures is based on the idea that the meaning of a sentence derives from what the speaker intends to communicate. Despite its foundational impact in linguistics and psycholinguistics, the universal application of this intention-based approach has been criticized. For example, the "intentionalist" (or "mentalist") view of pragmatics may not apply in cultures in which the "opacity of mind" ideology prevails (Robbins & Rumsey, 2008). In such cultures,

[1] Here we use the WEIRD/non-WEIRD distinction as a convenient characterization of the kinds of contexts in which research to date has been conducted, rather than a characterization of the contexts themselves. There is no underlying unity to "non-WEIRD" cultures, and we do not have any expectation that the development of pragmatic inference would be homogeneous across the range of cultures in the world. Our current study was intended to provide descriptive data on one culture — a first step in building data-driven expectations about the cross-cultural variability of pragmatic development.

people are reluctant to speculate about other people's intentions, and the mind is purported to be opaque and not easily readable. In Samoa, making pragmatic inferences has been described as less about understanding the speaker's intentions than looking at the social and material consequences of utterances. The role of invisible mental states is downplayed while that of the visible outcomes of speech acts is highlighted (Duranti, 1984, 2014). Similarly, Danziger (2006, 2010) has observed that the "opacity of mind" prevailing in the Mopan Maya culture of Belize undermines the very notions of intention and lie. Mopan Mayas do not seem to make any distinctions between a false sentence intending to deceive the listener and a false sentence whose falsehood is non-intentional. What really matters in their eyes is that a false sentence does not accurately depict the world, regardless of what the speaker's intention was.

Several investigators have also questioned whether the specific "maxims" of cooperative communication outlined by Grice (1975) are in operation consistently across cultures.[2] Drawing upon data collected in rural Madagascar, Ochs (1976) suggested that the maxim of informativeness is not used as extensively in other contexts as it is in Western culture. Similarly, Harris (1996) and Le Guen (2018) have pointed out that in rural Egypt and in Maya Yucatec culture, speakers are not generally expected to comply with truthfulness. On the contrary, Le Guen remarks that Yucatec Mayas' default expectation seems to be that lies and deception are pervasive.

When anthropologists and linguists question the purported universality of Gricean accounts, however, they are not claiming that the people they have studied on the field *never* comply with cooperative norms. The claim is rather that *in some situations* in which we would expect compliance with these norms in a Western context, no such compliance is to be found. Yet despite this general interest in pragmatic norms across cultures, and the importance of measuring the degree of compliance with Gricean accounts, relatively little work in the cross-cultural context has made use of new experimental paradigms designed to study pragmatic behaviors in the lab (e.g., Noveck & Reboul, 2008). In particular, experimental measurement might help researchers understand the degree to which patterns of reasoning are truly infelicitous vs. simply less common.

Pragmatic Development

The development of pragmatic abilities in childhood has been the focus of a deep literature. This work has examined a wide range of topics including the use of contextual, social, and discourse information (see e.g., Clark & Amaral, 2010) and the construction of common ground in word learning (for a review, see Tomasello, 2000). A

[2] Although this critique is posed in specifically Gricean language, we believe it applies equally to neo-Gricean accounts (e.g., Goodman & Frank, 2016).

particular focal point – with important implications for our study here – has been the question of the degree to which children make Gricean implicatures (e.g., Noveck, 2000; Papafragou & Tantalou, 2004; Barner et al., 2011; Frank & Goodman, 2014). One line of this work has examined performance in lexical scales such as quantifiers (e.g., "*some* of the cookies" implicates "*not all* of the cookies"; Noveck, 2000).[3] There is an emerging consensus that developmental issues in making such implicatures are related at least in part to knowledge of the individual scale members and their relationship to one another as alternatives (e.g., Barner et al., 2011; Horowitz et al., 2018).

In contrast, an alternative line of work has tried to measure children's performance in tasks where the relevant pragmatic implicature is created from contextual alternatives (e.g., Papafragou & Tantalou, 2004; Stiller et al., 2015). These tasks have the advantage of using situations that are easily accessible to children, offering the possibility of capturing developmental changes in the ability to make pragmatic inference. In Stiller et al. (2015), children were shown arrays containing three images, for example: [(1) a man], [(2) a man + glasses], [(3) a man + glasses + a hat]. They were then asked to help a puppet who said "My friend has glasses. Which one is my friend?". While the statement is literally true of both (2) and (3), on Gricean and other related accounts, an informative speaker would probably have said "hat" (or "hat and glasses") to describe (3). Thus, the puppet implicates pragmatically that (2) is his friend. In that study, children around 3.5 years old showed evidence of choosing (2) over the – presumably more interesting and salient – alternative (3).

Evidence from this study converges with data from a wide range of similar "ad-hoc" (contextually created) implicature tasks that show evidence of success around four years of age (Horowitz et al., 2018; Barner et al., 2011; Papafragou & Tantalou, 2004). While there has been some variation in the languages in which these tasks have been carried out (e.g., English, Greek), all of these studies have been conducted exclusively with Western populations, using convenience samples that typically reflect children recruited in WEIRD regions. Despite the simplicity of such tasks, and hence their suitability for translation across cultures and populations, little work has been done using them to investigate cross-population or cross-cultural differences in pragmatic inference.[4]

[3] An alternative perspective on implicature is the grammatical view, in which some – or all – implicatures are generated by the presence of a covert grammatical operator with the meaning "only", e.g. "only some of the cookies…" (Chierchia et al., 2012). This idea has received support in the literature on adults' scalar implicature (e.g., Franke & Bergen, 2020), but its application to children's pragmatic development is less accepted based on the successes of neo-Gricean models (e.g., Bohn et al., 2021).

[4] This pattern stands in contrast to work on quantification, which has made substantial progress cross-linguistically (Katsos et al., 2016).

The Current Study

The current study adapts the task described above from Stiller et al. (2015) to investigate cultural variation in pragmatic development, specifically in the Shipibo-Konibo (SK) people. The SK are an indigenous group living in the Peruvian Amazon, along the Ucayali River and its tributaries. They are mainly horticulturalists and fishermen (as well as occasional hunters), but are being increasingly integrated into the national Peruvian market economy. Although interactions with the Peruvian mestizo world – and even the Western world – are regular, SK culture remains very lively and still displays a strong identity. Although the SK language is well-studied from a linguistic perspective (e.g., Valenzuela, 2003), to our knowledge there is no specific evidence on SK pragmatics or related constructs (e.g., attitudes toward intention reading).

We conducted a variant of the ad-hoc pragmatic inference task described above with a group of SK children (4–11-year-olds). In general, SK children have a routine that is a mix between more traditional activities and educational activities. They tend to spend about 3 to 4 hours a day at school (every morning). Teaching at school is bilingual and this is how they are first exposed to Spanish language, but they do not master the basics of Spanish before early adolescence. When they are not at school, children spend their time playing with peers, without being monitored by adults. They are also quite involved in the daily tasks of their household (caring for younger siblings, gardening in the family *chacra*, fishing, cooking, etc.). Doing so, they learn a great deal of skills. As in many other indigenous cultures, learning occurs simply "by observing and pitching in" (Rogoff, 2014) and without any formal teaching (Lancy, 2016). As a result, SK children seem mature and autonomous compared to the average Western child.

In a pilot study, we tested SK children using the Stiller, Goodman, & Frank (2015) three-object paradigm described above. This paradigm proved to be difficult for young children, however (based on low performance even on control trials). As a consequence, in the present study, we used a simplified version of the paradigm that was designed for younger U.S. children and that involves computation of implicatures over two – instead of three – images (Yoon & Frank, 2019). Example stimuli are shown in Figure 1.

Figure 1. Example stimulus for a pragmatic inference (where the utterance would be "rice," with correct answer on the left) / control-double trial (where the utterance would be "fish," with correct answer on the right).

Methods

Participants

Children were recruited in two SK neighbourhoods of Yarinacocha, in the Pucallpa region of Peru, as well as in Bawanisho, a native community settled along the Ucayali River, 4 hours south of Pucallpa. Children were recruited either through their parents or through local schools. Data were collected from a total of 84 children, but 6 had no reliable age data associated and were excluded on this basis. The remaining 78 children were between the ages of 4 and 11 years old. Age of children was recorded as it was indicated by their DNI (Peruvian identity document). The 78 children in the final sample were split post-hoc into three approximately two-year age groups for descriptive and visualization purposes. Sample composition is shown in Table 1. Female children were more likely to participate because male children tended to be away from the village slightly more often.

Table 1. Sample composition

Age Group	Age (SD)	N	Male
4 – 6-year-olds	5.4 (.49)	11	3 (27%)
6 – 8-year-olds	7.1 (.50)	30	16 (53%)
8 – 11-year-olds	9.1 (.67)	37	14 (39%)

Stimuli

Our study had four trial types: *warm-up*, *control-single*, *pragmatic inference*, and *control-double*. Based on our earlier pilot study, we created a set of stimuli that were locally appropriate and that we believed would be easy for SK children to name (see Materials Availability, below).

Warm-up trials consisted of 4 consecutive trials where a participant needed to choose between 2 images. Although the pair belonged to the same superordinate category, they did not share any highly salient features. Warm-up trials were a *flower* vs. a *hat* (baseball cap), a *dog* vs. a *chicken,* a *chair* vs. a *ball,* and a *jaguar* vs. a *peccary* (local wild pig).

The main block of trials in the experiment consisted of two control-single trials, two pragmatic inference trials, and two control-double trials. Control-single trials were, like warm-up trials, choices between two different images, but this time more closely matched (images from the same basic-level category that differed on some property). The control-single trials were a *black-and-white kene* (fabric square) vs. a *colourful kene,* and a *gringo* couple (pair of Caucasian adults) vs. a *Shipibo-Konibo* couple (pair of SK adults).

In contrast, the base stimulus for both pragmatic inference and control-double trials was a pair of "containers" (e.g., *plate;* see Figure 1). Both containers shared one item (e.g., *rice* on the plate) and one had a unique item as well (e.g., *fish*). Items were *plates* with *fish* and *rice, motocarros* (vehicles) with *men* and *baskets, malocas* (traditional circular houses) with *trees* and *outhouses,* and *tables* with *plantains* and *aguaje* (moriche palm fruit). On pragmatic inference trials, the target word was the shared item (e.g., *rice*), with the intended referent being the container with only that item (e.g., the plate with *only rice*). On control-double trials, the target word was the unique feature (e.g., *fish*), with the intended referent being the container with both items.

We created four stimulus orders. Warm-up trials were given in a constant order, but trial type was counterbalanced for order in the six main trials. Target side was counterbalanced within each trial type. In addition, target item was counterbalanced across orders for the pragmatic inference and test trials (so that, e.g., *fish* was sometimes the shared item and sometimes the unique item). Similarly, the target word for warm-up and control-single trials was counterbalanced across orders.

Procedure

Children sat in front of the experimenter, whose hand was painted to look like a puppet. They were introduced to a fictional character called "Juanito" (the puppet) and were told that Juanito went for a walk and encountered different objects and people

on his way. Juanito would next ask children if they could help him locate these objects on the two images displayed in front of them. For example, the experimenter would say: "Juanito encountered a plate."[5] The puppet standing for Juanito would then ask: "this plate has rice; can you show me the plate?"[6] Children would have to point either to the [plate + rice] picture or to the [plate + rice + fish] picture. In this case, the pragmatically correct response was [plate + rice].

Children were first presented with four warm-up stimuli: i.e., stimuli very easy to discriminate (e.g., [jaguar] vs. [peccary]), to familiarize them with the task. They were then tested in a counterbalanced order on: two "control-single" trials (e.g., [coloured traditional fabric] vs. [black and white traditional fabric]); two "control-double" trials (e.g., [table + plantains] vs. [table + plantains + moriche palm fruit], after having been told that Juanito saw a table that has both plantains and moriche palm fruits); and two test trials (e.g., [table + plantains] vs. [table + plantains + moriche palm fruit], after having been told that Juanito saw a table that has plantains – implicating *only* plantains). The structure of the prompt was identical on all trials.

The instructions were translated into SK by a certified translator and the translation was subsequently revised by two SK bilinguals who are used to working with children; the whole experiment was performed in SK. Two sample videos are shared via Databrary (see Data Availability, below).

Results

Children's performance by age group across all trial types is shown in Figure 2. Across all age groups, children were at ceiling for warm-up and control-double trials, showing that they understood the task and were able to indicate the appropriate reference to the puppet. Both control-single and pragmatic inference trial performances were substantially lower, and close to chance except in the oldest age group.

---

[5] SK original version: "Juanitonin merai westiora rato."

[6] SK original version: "Nato rato riki arrozya; ¿Minki ea rato oinmati atipana?" The SK Research Assistant who performed the experiment introduced a slight procedural variation. Consistently with the procedure as just described, with some children, she used the puppet (i.e., she gestured with her painted hand as if the puppet was speaking) only to utter the final question: "can you show me the plate?" With some other children, on the other hand, the puppet was used both for the penultimate sentence "this plate has rice" and for the final question "can you show me the plate?" This slight procedural variation can be seen by comparing the two videos included in the Supplementary Materials. Importantly, what remained constant across children was that the first sentence ("Juanito encountered a plate") was always uttered by the experimenter and the last one ("can you show me the plate?") by Juanito.

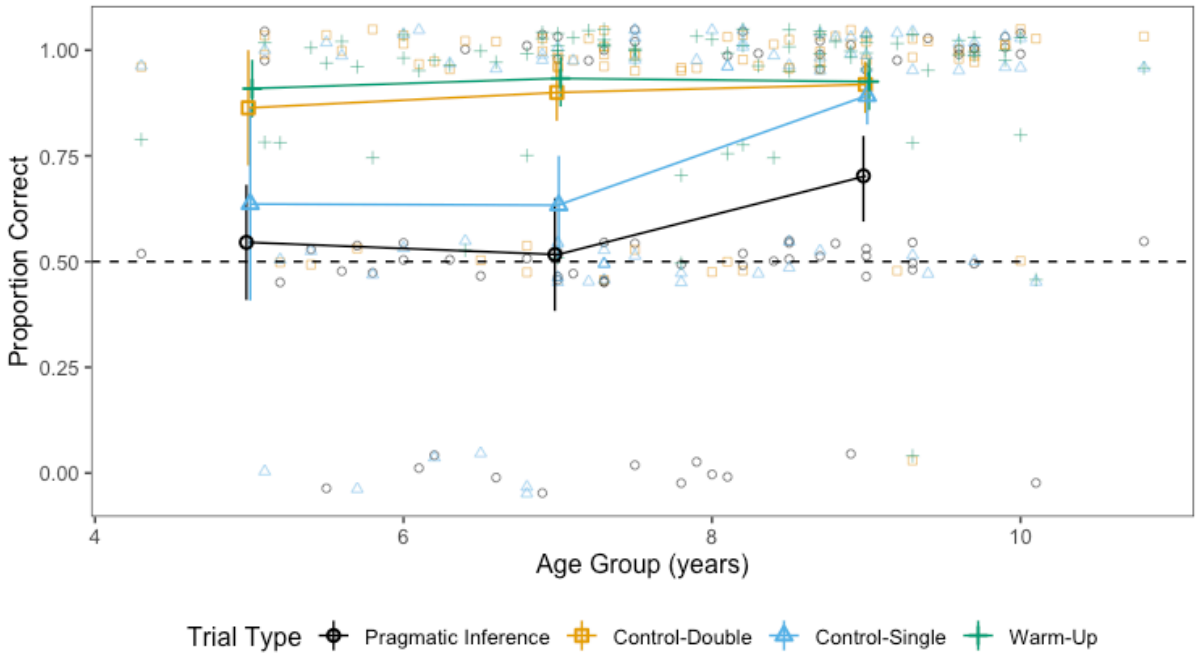


Figure 2. Proportion of correct (or pragmatically consistent, in the case of pragmatic inference trials) responses, plotted by age group. Error bars show 95% confidence intervals, computed by non-parametric bootstrap.

To investigate the strength of the evidence that children were above chance on the pragmatic trials, we computed default Bayesian *t*-tests using the BayesFactor package (Rouder et al., 2009) comparing children's mean responses to the null hypothesis of responding at chance. A first *t*-test revealed positive but relatively weak evidence for overall above-chance reporting across all children (BF_{10} = 4.25), but evidence was quite strong for 8–11-year-olds specifically (BF_{10} = 58.75).[7] These tests therefore support the conclusion of above chance pragmatic responding in the oldest children.

Children showed a similar pattern of performance for control-single trials, with BFs < 3 for the younger two groups, but very strong evidence for 8–11-year-olds specifically ($BF_{10} > 10^{10}$).[8] Why were children substantially weaker on control-single trials than control-double trials? We speculate that the items chosen ([Shipibo-Konibo cou-

[7] Note that the choice of this age group for follow-up analysis is post-hoc and reflects the division of the data into discrete age groups after data collection was complete.

[8] In all cases, qualitative conclusions were identical using frequentist *t*-tests (all BFs < 3 were non-significant at $p > .05$, and all BFs > 3 were significant at $p < .05$.

ple] vs. [Gringo couple] and [colourful traditional fabric] vs. [black and white traditional fabric]) must have been more difficult for children given that the trials are uncomplicated comparisons (but we do not have independent evidence on this question). But by design, our key comparison is control-double trials, which use the same materials as pragmatic inference trials but ask about the unique feature, rather than the repeated feature. In contrast to the control-single trials, the evidence from these trials was clear: only the oldest children were able to perform the pragmatic inference, but all children performed well on the control-double trials that used the same stimulus items (see Figure 3).

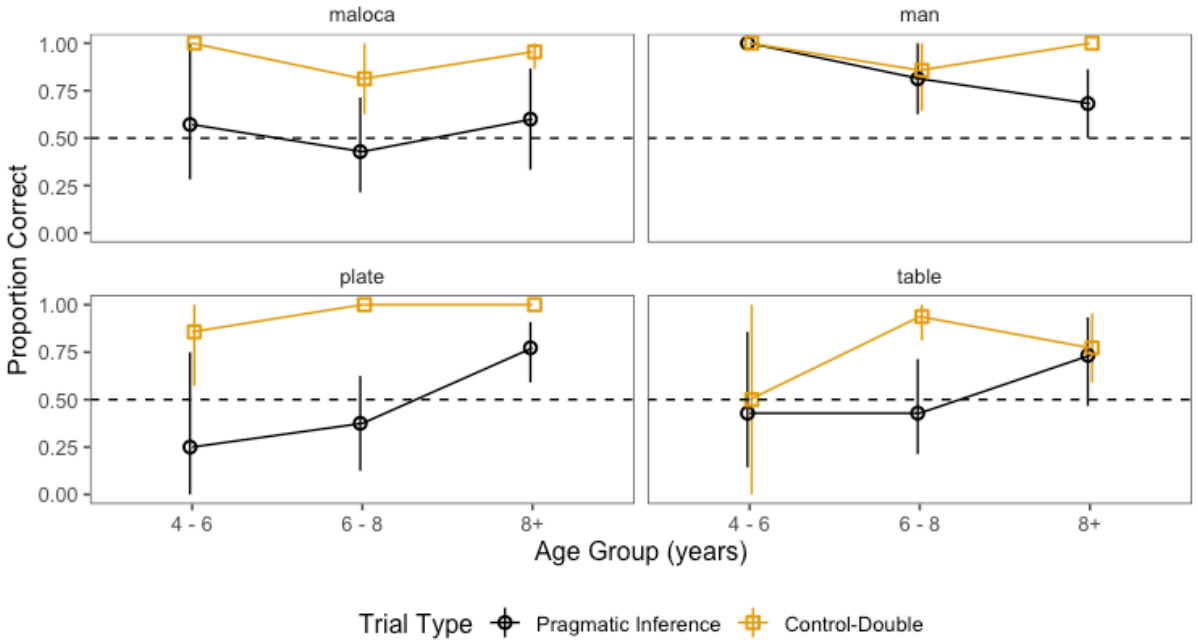


Figure 3. Proportion correct (or pragmatically consistent, in the case of pragmatic inference trials) responses, plotted by age group and experimental stimulus item. Error bars show 95% confidence intervals, computed by non-parametric bootstrap.

Discussion

How does pragmatic reasoning ability develop in children growing up in an indigenous Amazonian (and hence, non-WEIRD) culture? We used a simple ad-hoc implicature task adapted from previous work on pragmatic development to address this question in SK children. Although the younger children in our sample understood the task, they did not show the same patterns on the key pragmatic inference trials as has been observed in U.S. samples. In contrast, 8–11-year-olds showed relatively robust above-chance performance. Pragmatic inferences in our study were found substantially later in development relative to studies of children in the U.S. and Europe, where

three-year-olds show above chance performance in some tasks and four-year-olds are typically relatively accurate (e.g., Barner et al., 2011; Katsos & Bishop, 2011; Papafragou & Tantalou, 2004; Stiller et al., 2015; Yoon et al., 2018). Our findings nevertheless provide some new support for the idea that ad-hoc pragmatic inferences occur in a wide variety of cultural contexts.

The developmental differences we observed may relate to differences in children's language experiences. For example, SK children might experience fewer examples of pragmatic language use because more of their day-to-day language input is likely to come from peers rather than adults (Schneidman et al., 2012; Cristia et al., 2018). Young children overall tend to produce under-informative and egocentric language much more frequently than adults, even though they are in principle capable of reasoning about others' perspectives (see e.g., Nadig & Sedivy, 2002 for review). Such differences in input would result in differential familiarity with implicature and could create a more protracted developmental course. Many details in this hypothesis are underspecified, however. Even in U.S. contexts, the dependence of children's pragmatic inferencing on specifics of their language input is not completely understood, and this is even more true in the SK context.

The present design has several limitations that call for caution in the interpretation of our data and highlight the difficulty of cross-cultural research. First, in our paradigm, a fictional character (a puppet) was uttering sentences and asking children to compute implicatures. While U.S. children are comfortable with this type of setting, it must be stressed that interactions with fictional characters are virtually non-existent in SK culture and this feature likely rendered the paradigm more confusing. Performance in warm-up and control-double trials suggest that even younger children were able to answer simple questions, but they might still have struggled with the more complex and ambiguous test trials. Finally, the interpretation of our findings might differ depending on the correct account of implicature behaviour. It might be the case that ad-hoc implicatures are generated via a grammatical mechanism (following Chierchia, Fox, & Spector, 2012), and so our results might bear more directly on the availability of a grammatical operator (e.g., a covert "only") rather than – or in addition to – a pragmatic inference (Franke & Bergen, 2020).

Cross-cultural research should use a variety of paradigms and designs, not just one. Our results show that SK children's ability to compute ad-hoc implicatures is somewhat delayed as compared to U.S. and European children, but the generality of this result to other paradigms and methods of assessment is unknown. This question can only be answered by future research with both populations. As suggested by early critics of Grice, cross-cultural diversity in pragmatic inferences is never absolute: it is restricted to specific situations. The only way to test such subtle cross-cultural variations is to implement the richness of real-life pragmatic situations in a variety of experimental tasks.

References

Atran, S., & Medin, D. (2008). *The Native Mind and the Cultural Construction of Nature*. MIT Press.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition, 118*(1), 84–93.

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2021). How young children integrate information sources to infer the meaning of words. *Nature Human Behaviour, 5*(8), 1046-1054.

Bürkner, P. C. (2016). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28.

Chierchia, G., Fox, D., & Spector, B. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. Semantics: An international handbook of natural language meaning, 3, 2297-2332.

Clark, E. V., & Amaral, P. M. (2010). Children build on pragmatic information in language acquisition. *Language and Linguistics Compass,* 4(7), 445-457.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (in press). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development*. https://doi.org/10.1111/cdev.12974

Danziger, E. (2006). The Thought That Counts: The Interactional Consequences of Variation in Cultural Theories of Meaning. In N. Enfield & S. C. Levinson (Eds.), *Roots of Human Sociality. Culture, Cognition and Interaction* (pp. 259–278). Berg Publishers.

Danziger, E. (2010). On trying and lying: Cultural configurations of Grice's Maxim of Quality. *Intercultural Pragmatics, 7*(2), 199–219.

Duranti, A. (1984). Intentions, Self, and Local Theories of Meaning: Words and Social Action in a Samoan Context. *Center for Human Information Processing Technical Report, 122*, 1–22.

Duranti, A. (2014). *The anthropology of intentions: Language in a world of others*. Cambridge University Press.

Fortier, M., Wente, A. O., Fernández Flecha, M., & Gopnik, A. (Submitted). Abstract causal knowledge in Amazonia: Shipibo-Konibo horticulturalists learn multiple causality more readily than U.S. and Peruvian undergraduates.

Franke, M., & Bergen, L. (2020). Theory-driven statistical modeling for semantics and pragmatics: A case study on grammatically generated implicature readings. *Language, 96*(2), e77-e96.

Goodman, N., & Frank, M. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences, 20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics. Volume 3: Speech Acts* (pp. 41–58). Academic Press.

Grice, P. (1981). Presupposition and conversational implicature. In P. Cole (Ed.), *Radical Pragmatics* (pp. 183–198). Academic Press.

Harris, R. (1996). Truthfulness, conversational maxims and interaction in an Egyptian village. *Transactions of the Philological Society, 94*(1), 31–55. https://doi.org/10.1111/j.1467-968X.1996.tb01176.x

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61–135.

Horowitz, A. C., Schneider, R. M., & Frank, M. C. (2018). The Trouble With Quantifiers: Exploring Children's Deficits in Scalar Implicature. *Child Development*. https://doi.org/10.1111/cdev.13014

Katsos, N., & Bishop, D. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition, 120*(1), 67–81. https://doi.org/10.1016/j.cognition.2011.02.015

Katsos, N., Cummins, C., Ezeizabarrena, M. J., Gavarró, A., Kraljević, J. K., Hrzica, G., ... & Van Hout, A. (2016). Cross-linguistic patterns in the acquisition of quantifiers. Proceedings of the National Academy of Sciences, *113*(33), 9244-9249.

Lancy, D. (2016). Playing With Knives: The Socialization of Self-Initiated Learners. *Child Development, 87*(3), 654–665. https://doi.org/10.1111/cdev.12498

Le Guen, O. (2018). Managing epistemicity among the Yucatec Mayas (Mexico). In J.

Proust & M. Fortier (Eds.), *Metacognitive Diversity: An Interdisciplinary Approach*. Oxford University Press.

Lenaerts, M. (2004). *Anthropologie des Indiens Ashéninka d'Amazonie: Nos sœurs Manioc et l'étranger Jaguar*. L'Harmattan.

Lucas, C., Bridgers, S., Griffiths, T., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299. https://doi.org/10.1016/j.cognition.2013.12.010

Masuda, T., & Nisbett, R. (2001). Attending holistically versus analytically: comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology, 81*(5), 922–934.

Masuda, T., & Nisbett, R. (2006). Culture and change blindness. *Cognitive Science, 30*(2), 381–399. https://doi.org/10.1207/s15516709cog0000_63

Medin, D., & Bang, M. (2014). *Who's asking? Native science, western science, and science education*. MIT Press.

Medin, D., Ross, N., Atran, S., Cox, D., Coley, J., Proffitt, J., & Blok, S. (2006). Folkbiology of freshwater fish. *Cognition, 99*(3), 237–273.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science, 13*(4), 329-336.

Nisbett, R., & Masuda, T. (2003). Culture and point of view. *Proceedings of the National Academy of Sciences, 100*(19), 11163–11170. https://doi.org/10.1073/pnas.1934527100

Noveck, I. (2000). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition, 72*(2), 165–188.

Noveck, I. & Reboul, A. (2008). Experimental pragmatics: a Gricean turn in the study of language. *Trends in Cognitive Science, 12*(11), 425–431.

Ochs, E. (1976). The Universality of Conversational Postulates. *Language in Society, 5*(1), 67–80.

Papafragou, A., & Tantalou, N. (2004). Children's Computation of Implicatures. *Language Acquisition, 12*(1), 71–82. https://doi.org/10.1207/s15327817la1201_3

Robbins, J., & Rumsey, A. (2008). Introduction: Cultural and Linguistic Anthropology and the Opacity of Other Minds. *Anthropological Quarterly, 81*(2), 407–420.

Rogoff, B. (2014). Learning by Observing and Pitching In to Family and Community Endeavors: An Orientation. *Human Development, 57*(2–3), 69–81. https://doi.org/10.1159/000356757

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225-237

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: how important is directed speech? *Developmental Science, 15*(5), 659-673.

Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc Implicature in Preschool Children. *Language Learning and Development, 11*(2), 176–190. https://doi.org/10.1080/15475441.2014.927328

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics. Quarterly Publication of the International Pragmatics Association, 10*(4), 401-413.

Valenzuela, P. M. (2003). Transitivity in Shipibo-Konibo grammar (Publication #3095279). [Doctoral Dissertation: University of Oregon]. ProQuest Dissertations Publishing.

Walker, H. (2009). Baby hammocks and stone bowls: Urarina technologies of companionship and subjection. In F. Santos-Granero (Ed.), *The Occult Life of Things: Native Theories of Materiality and Personhood* (pp. 81-102). University of Arizona Press.

Washinawatok, K., Rasmussen, C., Bang, M., Medin, D., Woodring, J., Waxman, S., … Faber, L. (In Press). Children's Play with a Forest Diorama as a Window into Ecological Cognition. *Journal of Cognition and Development*. https://doi.org/10.1080/15248372.2017.1392306

Yoon, E., & Frank, M. (2019). The role of salience in young children's processing of ad hoc implicatures. *Journal of Experimental Child Psychology*, 186, 99–116. https://doi.org/10.1016/j.jecp.2019.04.008

Data, Code, and Materials Availability Statement

All data, code, and experimental materials are available at https://github.com/langcog/amazon_pragmatics. Example videos are available at

https://nyu.databrary.org/volume/691.

Ethics Statement

Our protocol received ethical approval from Pontificia Universidad Católica del Perú's Institutional Review Board. When recruited at school, consent for participation was collected from both the teachers and the parents; otherwise, only consent from the parents was collected. Although all children were eager to participate in the experiment, we could not test all of them, because some parents feared that we might be *pishtacos* (organ and blood thieves) and were thus reluctant to have their children involved.

Authorship and Contributorship Statement

MF, DK, MFF, and MCF designed research; MF collected data; DK and MCF analyzed data; MF and MCF drafted the initial paper; MF, DK, MFF, and MCF provided edits to the paper.

Acknowledgements

This paper is dedicated to the memory of Martin Fortier, who passed away in 2020. We wish to thank Margina Sampayo Mori for her valuable help in data collection. We are very grateful to all the SK teachers and community leaders who made data collection possible – in particular: Percy Cruz Laulate, Petronila Franco Márquez, Arturo Inuma, Sario Cruz, Roberto Matios, Alejandro Ochavano, and Santiago Matios. Many thanks to Manuel Bohn and Joëlle Proust for their helpful feedback on the manuscript. MF received funding from the France-Stanford Center for Interdisciplinary Studies, EHESS, and Fondation des Treilles. MCF was supported by NSF #1456077.

License

Uninversion error in English-speaking children's *wh*-questions: Blame it on the bigrams?

Ben Ambridge[1,2]
Stewart M. McCauley[3]
Colin Bannard[4,2]
Michelle Davis[1,2]
Thea Cameron-Faulkner[4,2]
Alison Gummery[2,5]
Anna Theakston[1,2]

University of Manchester, Division of Psychology, Communication and Human Neuroscience[1]
ESRC International Centre for Language and Communicative Development (LuCiD)[2]
University of Iowa, Department of Communication Sciences and Disorders [3]
University of Manchester, Linguistics and English Language[4]
University of Liverpool, Psychology[5]

Abstract: The aim of the present study was to investigate whether and how English-speaking children's uninversion errors with wh-questions (e.g., **Who he can draw; c.f., Who can he draw?*) are influenced by the surface frequency of individual bigrams and trigrams in the input, as predicted by input-based approaches. Production methods were used to elicit nonsubject *wh*-questions from 67 children aged 3;1 to 4;8 (*M*=4;0, *SD*=4 months). No support was found for the preregistered prediction that children will produce more uninversion errors when those errors incorporate – in the Bigram 3 position – high-frequency bigrams from uninverted structures (e.g., **Who he* ***can draw****?*) than lower-frequency bigrams from uninverted structures (e.g., **Who he* ***can name?***). Importantly, when testing this prediction, all other bigrams and unigrams (i.e., single words) are either identical (e.g., *Who+he, he+can, he, can*) or closely matched for frequency (e.g., *draw* and *name* [as verbs] are of approximately equal corpus frequency). However, a non-preregistered exploratory analysis found a facilitatory effect on correct-question production of the frequency of the second and third bigrams from inverted structures (e.g., *can he…he draw*), even after controlling for unigram frequency. This analysis also found that rates of uninversion error (e.g., **Who he can draw?*) were higher when the first uninverted bigram (e.g., *Who he…*) is of higher frequency in the input. We conclude that while input-based accounts are correct to highlight the importance of n-gram input frequencies on rates of correct production versus uninversion error, it is unclear on current evidence which n-grams are driving errors and why. In particular, the special emphasis placed by some such accounts on n-grams at the left-edge of the utterance (e.g, *Who can…*) may be unwarranted.

Keywords: *wh*-questions, elicited production, elicited imitation, frequency.

Corresponding author(s): Ben Ambridge, Psychology, Communication and Human Neuroscience, Coupland Building 1, University of Manchester, Manchester, UK, M15 6FH.

ORCID ID(s): https://orcid.org/0000-0003-2389-8477

Citation: Ambridge, B., McCauley, S., Bannard, C., Davis, M., Cameron-Faulkner, C., Gummery, A., Theakston, A. Uninversion error in English-speaking children's *wh*-questions: Blame it on the bigrams?. *Language Development Research*, *3*(1), 121–155. https://doi.org/10.34842/2023.641

Introduction

Wh-questions occupy a special place in language development research, since they are the only commonly used sentence-level construction for which English-speaking children regularly produce word-order errors; specifically, uninversion (or non-inversion) errors[1] such as **Who he can draw?* (cf., *Who can he draw?*). Early interest in these errors (Bellugi, 1971; Hurford, 1975; Kuczaj, 1976; Tyack & Ingram, 1977; Maratsos & Kuczaj, 1978; Labov & Labov, 1978; Kuczaj & Brannick, 1979; Bloom, Merkin & Wooten, 1982; Erreich, 1984) was sparked by the fact that they appear to reflect children's failure to apply a particular form of syntactic movement (I-to-C movement, or *subject-auxiliary inversion*; e.g., *Who he can draw?* → *Who can he draw?*) having already moved the *wh-* word from its corresponding position in declarative utterances (e.g., *He can draw who* → *Who he can draw*).

Subsequent accounts developed in this movement- or rule-based framework have sought to explain why children fail to apply this movement rule to particular *wh*-words (DeVilliers, 1991; Valian, Lasser & Mandelbaum, 1992; Pozzan & Valian, 2017), auxiliaries (Santelmann, Berk, Austin, Somashekar, & Lust, 2002; Hattori, 2003; Westergaard, 2009), or both (Stromswold, 1990, 1995; Valian & Casey, 2003).

In contrast, accounts developed in a usage-based (or "constructivist") framework have sought to explain these errors (sometimes referred to as "non-target-consistent" or simply "ungrammatical" questions) in terms of properties of the input. We term these accounts "input-based" because – although all accounts must of course posit *some* role for the input – such accounts claim that children are learning the structure of questions directly from the input, rather than merely using the input to trigger rules or parameters (e.g., wh-movement; I-to-C movement' subject-auxiliary inversion). That said, as we will see shortly, different varieties of input-based account potentially make subtly different predictions regarding frequency effects in question production.

Consistent with input-based approaches (in the broad sense), several studies have shown that children are less likely to produce uninversion errors (e.g., **Who he can draw?*) when lexical strings that appear in the correct form – particular *wh*-word+auxiliary combinations, such as *who can* – are frequent in the input (e.g., Rowland & Pine, 2000; Rowland, Pine, Lieven & Theakston, 2003, 2005; Dabrowka & Lieven, 2005; Ambridge, Rowland, Theakston & Tomasello, 2006; Rowland, 2007; Ambridge & Rowland, 2009). McCauley, Bannard, Theakston, Davis, Cameron-Faulkner and Ambridge (2021) also showed that children are more likely to produce uninversion errors (e.g., **Who he can draw*) when lexical strings that appear in the errorful form are frequent in the input (e.g., *he can* is considerably more frequent than *can he*). Of course, these findings do not demonstrate the *absence* of a syntactic subject-auxiliary inversion rule. What they do suggest however, is that – at the very least – the output of such a rule is filtered through a production mechanism that is sensitive to the input frequency of multiword strings (e.g., bigrams and trigrams; collectively n-grams); a

[1] Technically, "uninversion" incorrectly implies that the erroneous questions started out as inverted, and were then "uninverted". However, because the term is more widespread in the literature than the slightly more cumbersome term "non-inversion" errors, we use it (and "uninverted") throughout.

mechanism that can both cause errors and protect against them (Ambridge, Rowland, Theakston & Kidd, 2015).

The aim of the present study was to conduct a particularly tightly controlled investigation of input-based accounts of uninversion errors by investigating the effect of the input frequency of the third bigram in uninverted questions, while holding constant the frequency of all other bigrams (e.g., **Who he **can draw?*** [high-frequency] vs. **Who he **can name?*** [low-frequency]). This constitutes something of a departure from most studies in this domain (McCauley et al., 2021, excepted), which have generally focused on n-grams towards the left edge of the utterance and – in the main – on n-grams that appear solely or mainly in questions (e.g., *who can* or *who can he*), and that therefore support correct-question formation, rather than causing uninversion errors. Having conducted a preregistered test of this prediction, we then go on to conduct exploratory analyses in which we investigate in a more open-ended fashion input-frequency effects for other n-grams; again, both n-grams from inverted structures (mainly questions) that protect against inversion errors, and n-grams from uninverted structures (mainly declaratives) that cause inversion errors.

The starting point for the present study is the corpus study of McCauley et al. (2021) which, in turn, was inspired by studies showing faster processing and/or fewer production errors for higher frequency n-grams, for both adults (e.g., Liberman, 1963; Krug, 1998; Bybee & Scheibman, 1999; Jurafsky, Bell, Gregory, & Raymond, 2001; Sosa & MacFarlane, 2002; McDonald & Shillcock, 2003; Pluymaekers, Ernestus, & Baayen, 2005; Bannard, 2006; Arnon & Snider, 2010; Tremblay & Baayen, 2010; Siyanova-Chanturia, Conklin, and van Heuven, 2011; Janssen & Barber, 2012; Hernández, Costa & Arnon, 2016; Arnon, McCauley & Christiansen, 2017) and children (Bannard & Matthews, 2008; Arnon & Clark, 2011; Havron & Arnon, 2021; Skarabela, Ota, O'Connor & Arnon, 2021; Kueser & Leonard, 2020).

In an analysis of 12 spontaneous speech corpora from the English-speaking portion of CHILDES (MacWhinney, 2000), McCauley et al. (2021) showed that the frequency of children's uninversion errors versus correct questions (e.g., **What you are doing there* vs *What are you doing there?*) was (a) negatively related to the input frequency of the third and fourth bigram in the correct, inverted question (e.g., *you doing; doing there*) and (b) positively related to the input frequency of the second, third and fourth bigram in the errorful, uninverted question (e.g., *you are, are doing, doing there*). To clarify, the reason that children were hearing "uninverted" bigrams such as *you are, are doing* and *doing there* was NOT because their caregivers were producing uninversion errors; they were not. Rather, children were hearing these "uninverted" bigrams as part of declarative sentences (e.g., ***You are** happy; They **are doing** it*), complement clauses (*I wonder what he's **doing there***), including those used for reported questions (e.g., *He asked whether **you are doing** it*), and so on. That is, even though these children were easily capable of distinguishing questions from declaratives and other non-questions, high-frequency uninverted n-grams heard in the context of declaratives constituted "lures" towards uninversion errors in question production; albeit lures that children could resist when the target inverted n-grams (i.e., those heard in the context of questions) were of sufficiently high input frequency. These findings are summarized in Figure 1 (reproduced from McCauley et al., 2021, under the terms of the Creative Commons CC-BY license, which permits unrestricted use).

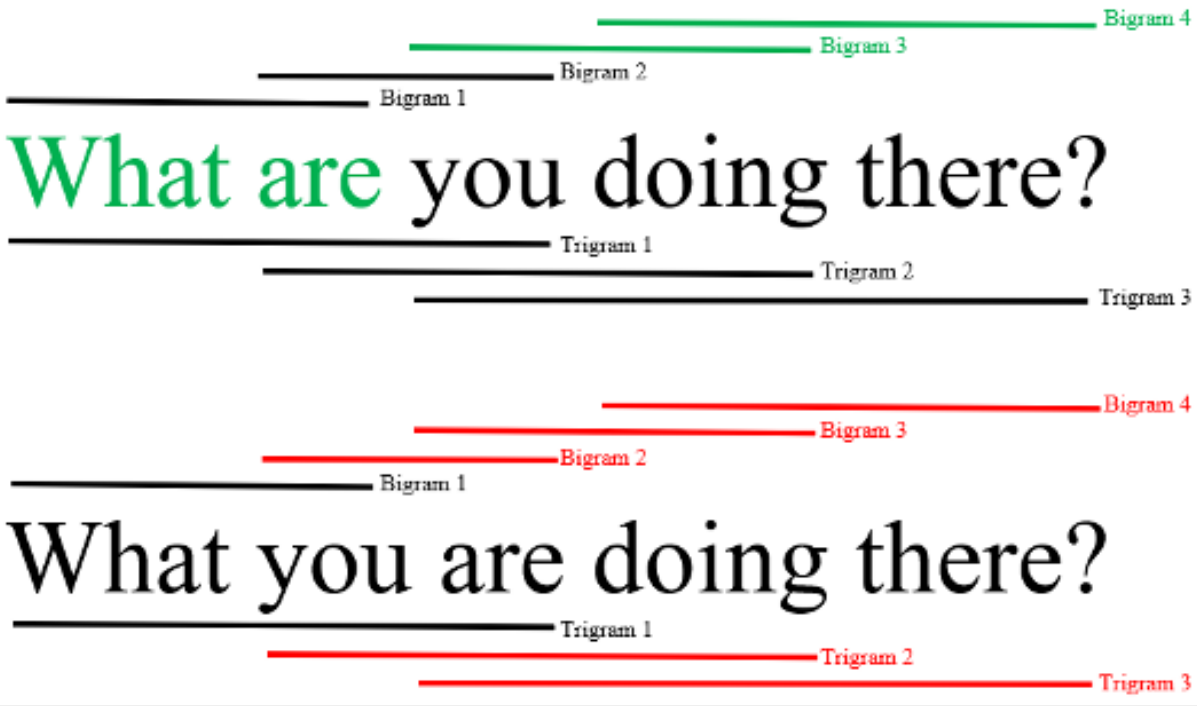


Figure 1. ***Summary of the findings of the corpus study of McCauley et al. (2021). Unigrams (individual words), bigrams, and trigrams for the correct, inverted (top) and corresponding errorful, uninverted (bottom) forms of the example question What are you doing there? N-grams excluded from the final statistical model are shown in black. N-grams retained in the final statistical model are shown as green/red words (unigrams) and green/red lines (bigrams and trigrams).***

Indeed, there is precedent for McCauley et al's (2021) finding that high frequency input strings from one sentence type (here, mainly declaratives) can constitute "lures" towards errors for a different sentence type (here, questions). For example, in Norwegian (like many V2 languages), the negation marker appears after the verb in main clauses (e.g., *We* ***read not*** *Icelandic sagas every night*) but before the verb in embedded clauses (e.g., *The teacher knows that we* ***not read*** *Icelandic sagas every night*). Children learning Norwegian often make errors when attempting to produce embedded clauses (e.g., **The teacher knows that we* ***read not*** *Icelandic sagas every night*), by inappropriately generalizing on the basis of high-frequency combinations with main-clause word order, here ***read+not*** (Westergaard & Bentzen, 2007; Ringstad & Kush, 2021; see also Waldmann, 2012, for a similar finding in Swedish). This is analogous to McCauley et al's (2021) finding that high-frequency n-grams from (mainly) declaratives (e.g., ***you are***) lead to uninversion errors in question formation (e.g., **What* ***you are*** *doing?*).

Perhaps surprisingly, unlike previous studies (e.g., Rowland & Pine, 2000; Ambridge et al., 2006; Ambridge & Rowland, 2009) McCauley et al. (2021) found no significant frequency effect of the first inverted bigram, which – for *wh*-questions – is always a *wh*-word+auxiliary combination (e.g., *What are*; *What is*; *Why is*; *Who can* etc…). However, this may be a consequence of the unusually strict analysis used by McCauley et al. (2021), under which bigram frequency effects were investigated only after controlling for frequency effects at the level of each individual lexical item (or "unigram").

Indeed, significant unigram frequency effects were observed for the first two inverted positions (e.g., *What; are*). Thus, we cannot conclude that the frequency of the first inverted bigram (*wh*-word+auxiliary) has no effect; only that we cannot detect an effect of the *wh*-word+auxiliary combination above and beyond frequency effects observed for the *wh*-word and auxiliary individually.

The aim of the present study was to conduct an experimental test of a prediction that follows from the study of McCauley et al. (2021), and from the more general claim of (at least some) input-based approaches, that learners retain, and are influenced by, individual lexical strings even when they have formed more abstract representations too (e.g., Langacker, 1998; Abbot-Smith & Tomasello, 2006; Goldberg, 2006; Ambridge 2020a, 2020b). That prediction, as preregistered at https://osf.io/tbmu4 prior to data collection (registration DOI: https://doi.org/10.17605/OSF.IO/TBMU4), was as follows:

> *Wh-* questions are more likely to be produced without subject-auxiliary inversion when the multiword sequences making up the errorful, non-inverted form of a child's target question are of a higher frequency. That is, subjects will make more uninversion errors with questions in the high frequency condition (where the frequency of "can go" in the uninverted form of the question "where can he go?" is high) than in the low frequency condition (where "can play" in the uninverted form of the question "where can he play?" is of a lower frequency, relative to "can go," while the correctly form[ed] questions are matched for the frequency of all trigrams, bigrams, and unigrams).

In order to have control over the target questions that children were attempting to produce, it was necessary to use an elicited-production methodology, in which the experimenter produced the target *wh*-word, auxiliary, subject and verb, but in uninverted order (as per Ambridge et al., 2006, 2008; Ambridge & Rowland, 2009). The method can be summarized as follows (again, quoting from our preregistration document):

> The experiment is couched in terms of a "jigsaw puzzle" game where the child is asking questions to a toy dog…In each trial, the child is prompted to produce a question by the experimenter by showing them an image consisting of one or more "jigsaw puzzle" pieces. Slots for missing jigsaw pieces are apparent in this image, and conceal some aspect of the target question. For instance, the missing jigsaw pieces may be hiding a ball in the case of a trial involving the target question "What is she holding?" The experimenter then attempts to elicit the target question by saying "I wonder what she's holding? Let's ask the dog what she is holding!" When the child asks the question, the missing jigsaw pieces are then filled in to reveal the ball (in the case of this example trial). The child then hears an audio recording (meant to be the dog's voice) answering the question. In this case, "A ball!"

Before setting out the present study in detail, it is important to clarify that not all "input-based" accounts of question acquisition would necessarily share the prediction set out above (or the non-preregistered effects that we uncovered in subsequent, exploratory analyses). For example, Rowland and Pine (2000), Dabrowska and Lieven

(2005), Ambridge et al. (2006) and Ambridge & Rowland (2009) all posit that children, certainly by age 3-4, form slot-and-frame question schemas such as *What are [THING] [PROCESS]?* Because these are informal, verbal accounts (as opposed to formal mathematical or computational models) they do not yield precise quantitative predictions. But one possible interpretation of these accounts – and quite possibly the dominant one in the literature – is that only the "frame" (e.g., *What+are*) is fixed, with the "slots" [THING] [PROCESS] free. Consequently, such accounts arguably predict that the frequency of words or combinations in the slot positions will not affect rates of correct production versus uninversion error.

In the present study, however, we test a different, more radically-exemplar-based type of input-based account (e.g., Langacker, 1998; Abbot-Smith & Tomasello, 2006; Goldberg, 2006; Ambridge 2020a, 2020b; McCauley et al., 2021).,which assumes that whether or not children form, in some sense, "free slots", they remain sensitive to the frequency of the individual n-gram combinations in the exemplars that gave rise to those slots.

Method

Ethics

Ethics approval was obtained from the University of Liverpool Research Ethics Committee prior to recruitment. Children's caregivers gave informed written consent and children gave verbal consent.

Participants

Our preregistration specified a minimum of 60 (providing 90% power) and a maximum of 70 participants, chosen on the basis of a power analysis calculation conducted using the "simr" R package (for details see https://osf.io/74urw/) assuming α=0.05. The simulation data were based on a small pilot study (N=12), but were adjusted to assume a small effect size for our primary manipulation (d=0.2), since no such effect was present in the pilot data. All children were native learners of English, with no known language impairments, and received stickers for their participation.

Given that our primary manipulation compares rates of uninversion errors within matched question pairs (e.g., *Who can he draw?* vs *Who can he name?*), it was important to ensure that we recruited a sufficient number of participants who produced scoreable responses (correct questions or uninversion errors with the target lexical items) for *both* questions in a given pair. Our preregistration therefore stipulated that "We will retain data only from children who produce scorable responses (correct question or noninversion errors) for a minimum of three high+low frequency pairs. Any excluded participants will be replaced in order to ensure our target sample size of 60". Of the 113 children who began the study, 46 were excluded and replaced on this basis, for a final sample size of N=67. Although a drop-out rate of 40% may seem high, it partly reflects the fact that – due to our focus on particular n-grams – it was necessary to exclude otherwise-scorable questions that included perfectly reasonable substitutions (e.g., *Who can the man draw?* for *Who can he draw?*). The final sample ranged in age from 3;1 to 4;8 with a mean of 4;0 (SD=4 months).

Design and Materials

The primary aim of the study was to test the prediction that participants will produce more uninversion errors when those errors incorporate – in the Bigram 3 position – high-frequency (e.g., **Who he **can draw**?*) rather than lower-frequency bigrams from uninverted structures (e.g., *Who he **can name**?*). Recall that when testing this prediction, all other bigrams and unigrams (i.e., single words) are either identical (e.g., *Who+he, he+can, he, can*) or closely matched for frequency (e.g., *draw* and *name* [as verbs] are of approximately equal corpus frequency). Using n-gram frequencies from the child-directed portion of the entire CHILDES database for both UK and US English (MacWhinney, 2000) – a total of 3,436,333 utterances (not selected or coded for sentence type) – we created eight question pairs that met this criterion (see Table 1). It is important to make clear at this point that, as explained in further detail below, the experimenter's prompt sentences in fact included uninverted questions; albeit grammatically acceptable ones that constitute reported speech (e.g., *I wonder who he can draw*). Thus in order to produce a well-formed question, the child has to "invert" the experimenter's prompt question.

With regard to the frequency of the n-grams from inverted structures, the "high" and "low" frequency questions in each pair (defined with regard to Uninverted Bigram 3) were perfectly matched for Bigram 1 and Bigram 2 (since the first three words were identical) and approximately matched for Bigram 3 (and likewise for the corresponding trigrams)[2]. For example, consider the question pair *Who can he draw? / Who can he name?*, which yield the uninversion errors **Who he can draw? / *Who he can name?*. With regard to the frequency of the n-grams from inverted structures, the two are identical with respect to Bigram 1 (*who can*) and Bigram 2 (*can he*), and closely matched for Bigram 3 (*he draw* = 15 occurrences in the corpus; *he name* = 12 occurrences). The high and low frequency questions of each pair were closely matched with regard to the frequency of n-grams from inverted structures, in order to allow for a specific and highly controlled investigation of the "lure" effects of n-grams from uninverted structures. That is, the experiment asks: "Even though the correct forms *Who can he draw? and Who can he name* are **equally probable statistically,** are uninversion errors more common for the first than the second, since the "lure" bigram *can draw* (**Who he **can draw**?*) is more frequent than the "lure" bigram *can name* (**Who he can name?*)?

With regard to the frequency of the n-grams from uninverted structures, the "high" and "low" frequency (defined with regard to Uninverted Bigram 3) questions in each pair were again perfectly matched for Bigram 1 and Bigram 2 (since the first three words were identical), but **mismatched** as far as possible for Bigram 3 (and likewise for the corresponding trigrams), such that the high-frequency bigram was, in each case, at least 10 times as frequent as the low-frequency bigram. For example,

[2] In these types of circumstances, researchers often report a significance test to show that the "matched" items did not "differ significantly" on the value in question (here, frequency). However, this is not appropriate since such tests are properly used to generalize instances made from a sample to a wider population, and cannot meaningfully be used to draw conclusions about an entire population; here, of test items (Sassenhagen & Alday, 2016).

Table 1. ***Stimulus pairs and n-gram frequencies.***

Target	**Cond.**	**Inverted Trigram1** *who can*	**Inverted Trigram2** *can he draw*	**Inverted Bigram1** *who can*	**Inverted Bigram2** *can he*	**Inverted Bigram3** *he draw*
Who can he draw?	High	6	0	258	850	15
Who can he name?	Low	6	2	258	850	12
What can he eat?	High	42	10	1686	850	323
What can he need?	Low	42	0	1686	850	243
What can he hear?	High	42	13	1686	850	34
What can he mean?	Low	42	2	1686	850	19
Where is Daddy sitting?	High	209	0	34260	578	2
Where is Daddy singing?	Low	209	0	34260	578	2
What can it hold?	High	14	1	1686	226	56
What can it cause?	Low	14	0	1686	226	59
What could it see?	High	24	0	257	158	65
What could it want?	Low	24	0	257	158	50
Why is Daddy hiding?	High	9	0	2469	578	0
Why is Daddy building?	Low	9	0	2469	578	0
What is it wearing?	High	3012	0	87230	19083	0
What is it kissing?	Low	3012	0	87230	19083	0

Target	**Cond.**	**Uninverted Trigram1** *who he can*	**Uninverted Trigram2** *he can*	**Uninverted Bigram1** *who he*	**Uninverted Bigram2** *he can*	**Uninverted Bigram3** *can draw*
Who can he draw?	High	0	6	117	3260	316
Who can he name?	Low	0	1	117	3260	16
What can he eat?	High	33	75	2924	3260	817
What can he need?	Low	33	0	2924	3260	2
What can he hear?	High	33	60	2924	3260	1060
What can he mean?	Low	33	0	2924	3260	9
Where is Daddy sitting?	High	3	1	90	198	579
Where is Daddy singing?	Low	3	0	90	198	57
What can it hold?	High	11	1	2499	954	313
What can it cause?	Low	11	0	2499	954	8
What could it see?	High	7	1	2499	719	313
What could it want?	Low	7	0	2499	719	1
Why is Daddy hiding?	High	1	0	14	198	335
Why is Daddy building?	Low	1	0	14	198	31
What is it wearing?	High	359	0	2499	8081	352
What is it kissing?	Low	359	0	2499	8081	27

Target	**Cond.**	**Unigram1** ***who***	**Unigram2** ***can***	**Unigram3** ***he***	**Unigram4** ***draw***	**Dog's answer**
Who can he draw?	High	41853	102758	212458	5466	His mum!
Who can he name?	Low	41853	102758	212458	6296	His new puppy!
What can he eat?	High	269958	102758	212458	22551	His breakfast!
What can he need?	Low	269958	102758	212458	23302	A new pair of shoes!
What can he hear?	High	269958	102758	212458	7725	A Bird!
What can he mean?	Low	269958	102758	212458	8214	That he is hungry!
Where is Daddy sitting?	High	76055	348124	14295	4212	In the kitchen!
Where is Daddy singing?	Low	76055	348124	14295	3587	In the garden!
What can it hold?	High	269958	102758	260253	8677	A toy
What can it cause?	Low	269958	102758	260253	18436	An accident
What could it see?	High	269958	18299	260253	66313	A mouse!
What could it want?	Low	269958	18299	260253	94362	Cat food!
Why is Daddy hiding?	High	29443	348124	14295	2073	He's playing hide and seek
Why is Daddy building?	Low	29443	348124	14295	1568	He's playing with LEGO
What is it wearing?	High	269958	348124	260253	1550	A sweater!
What is it kissing?	Low	269958	348124	260253	758	Its mum!

considering again the question pair *Who can he draw? / Who can he name?*, with regard to the frequency of the n-grams from uninverted structures, the two are identical with respect to Bigram 1 (*who he*) and Bigram 2 (*he can*), while Bigram 3 is approximately 20 times more frequent for the high-frequency version (*can draw* = 316) than the low-frequency version (*can name* = 12).

In response to presentations of this and previous work, colleagues have often expressed surprise that children hear "uninverted" bigrams (e.g., *who he, he can, can draw*) in the input at all, given that parents and other adults produce few, if any, uninversion errors. It is therefore important to remind the reader that children heard these "uninverted" (with respect to questions) bigrams as part of declarative sentences, including those used for reported speech (e.g., *I wonder* ***who he*** *means;* ***He can*** *do it; You* ***can draw*** *it*). The hypothesis under investigation (which enjoys preliminary support from the study of McCauley et al., 2021) is that, despite having been heard solely in declaratives, these n-grams constitute "lures" towards uninversion errors in question production.

Procedure

The experimenter began the (single) session with the following general instructions:

> Hi, my name is [xxx] and we're going to play a special game with this talking dog. It's a girl dog, and her name is Fifi *[note: this was to ensure that "he" when used in the target questions could not refer to the dog]*. We've got some jigsaws here *[Show Warm-up 1a]* but, uh oh, the jigsaws are all missing some pieces so we can't see what's happening. Luckily, Fifi has got the missing pieces, so we can

> ask her what's happening. Then she'll put in the missing pieces. Don't worry, I'm going to help you by telling you what to ask Fifi.

Showing the first warm-up picture onscreen (see Figure 2a; presented via an Open Sesame script; https://osdoc.cogsci.nl), the experimenter continued:

> So, in this first one, we've got a girl called Sarah. Do you know any girls called Sarah? OK, anyway, so here's Sarah. In this jigsaw, she's carrying something. I wonder what Sarah is carrying. Let's ask the dog what Sarah is carrying. Copy me. Say "What is Sarah carrying?" [Note: in the first two warm-up trials, the experimenter invited the child to copy her question verbatim].

After the child's response (*What is Sarah carrying?*), the experimenter activated the "talking dog" toy to have it produce a pre-recorded answer (here, *a book*). At the same time, the missing pieces of the jigsaw appeared onscreen (see Figure 2b).

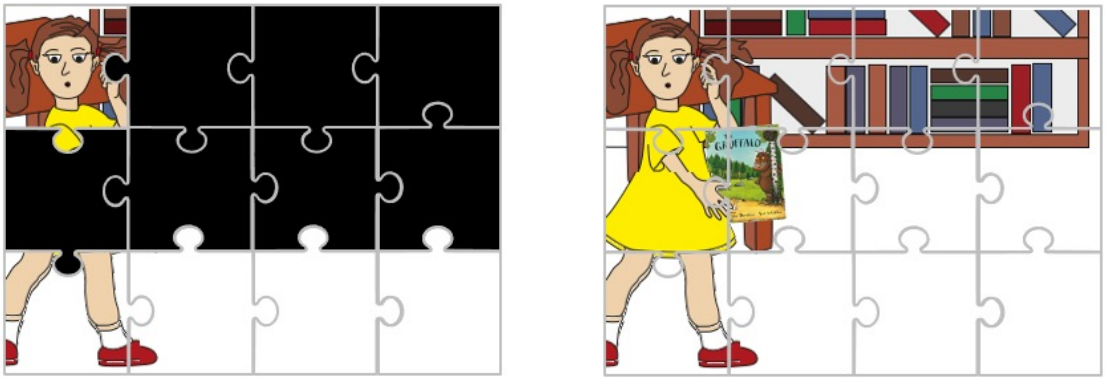


Figure 2. ***Before (2a) and After (2b) pictures shown to children for the first warm-up trial: Q: What is Sarah carrying? A: A book!***

A second warm-up trial (*What is Sarah giving?*) proceeded in the same way, with the child copying the experimenter's question verbatim. For these first two warm-up trials only, the experimenter corrected children who did not produce the target question. For the third warm-up trial, the experimenter announced:

> Now, this time, you're not going to copy me. Instead, I'll just tell you what to ask and you ask it OK? Don't worry, I'll still tell you what to ask. So here's Sarah again. In this jigsaw, she's throwing something. I wonder what Sarah is throwing. Let's ask the dog what Sarah is throwing. You ask the dog what Sarah is throwing.

Note that, for this warm-up trial, and the final, fourth, warm-up trial, the experimenter used indirect/reported speech to present the target question string (grammatically) in uninverted order (*what Sarah is throwing; what Sarah is pushing*). Although the experimenter was careful to always use declarative intonation (i.e., not question intonation), it is important to acknowledge that this method to some extent primes children to produce uninverted questions, both at the abstract level (e.g., *[wh-word] [SUBJECT] [BE] [VERB*]?) and the lexical level (e.g., Savage, Lieven, Theakston, &

Tomasello, 2003; Huttenlocher, Vasilyeva, & Shimpi, 2004; Bencini & Valian, 2008; Rowland, Chang, Ambridge, Pine, & Lieven, 2012). Whether or not this constitutes a confound that potentially invalidates any pattern of uninverted forms found in the data is a question to which we return in the Discussion.

Thus (amongst other possible responses) children could repeat the sequence produced by the experimenter verbatim, yielding an uninversion error, or "invert" the experimenter's question, yielding a correct response. From the third warm-up trial onwards, the experimenter did not correct children's questions, providing only general encouragement. After the final warm-up trial, the experimenter said "Brilliant! OK, now let's try some more pictures with different people in", and proceeded to the 16 test trials, which worked in the same way as the final two warm-up trials. The prompts for the test trials can be found in Appendix 3. Note that while, for warm-up trials, the SUBJECT was always *Sarah*, for the test trials, the SUBJECT was always *he*, *Daddy* or *it*.

In order to sufficiently separate the presentation of the high- and low-frequency (with regard to Uninverted Bigram 3) members of each question pair, the 16 trials were divided into two blocks of 8, presented consecutively. For each participant, two pseudo-randomized lists were created such that if the high-frequency member of a particular question pair appeared in Block 1 (N=4) the low-frequency member of that pair appeared in Block 2 (N=4), and vice versa for the remaining 4 pairs. Within each block, the order of presentation was fully randomized.

Results

We first present the results of our main pre-registered analysis before presenting a number of exploratory analyses designed to investigate the role of the frequency of particular n-grams.

Main, pre-registered analysis

The pre-registered analysis was designed to test the prediction that participants will produce more uninversion errors when those errors incorporate – in the Bigram 3 position – high-frequency bigrams from uninverted structures (e.g., **Who he **can draw**?*) than lower-frequency bigrams from uninverted structures (e.g., **Who he **can name?***). Importantly, when testing this prediction, all other bigrams and unigrams (i.e., single words) are either identical (e.g., *Who+he, he+can, he, can*) or closely matched for frequency (e.g., *draw* and *name* [as verbs] are of approximately equal corpus frequency). Note that, for this analysis, the frequency of the crucial bigram is treated as a categorical predictor (Condition: high/low) since it is manipulated within each otherwise-closely-matched target question pair. Exploratory analyses presented below investigate continuous frequency effects.

Thus, the following pre-registered mixed-effects models syntax[3], for the lme4 package (Bates, Mächler, Bolker & Walker, 2015) in the R environment (R Core Team,

[3] In fact, this model is not quite optimal given the study design, as it fails to take account the fact that TargetSentence is nested inside sentence pair (the pair of sentences matched for n-gram frequency other than the target one). In fact, a model including random slopes for both TargetSentence and

2022), was designed to test the hypothesis of a main effect of condition (high/low frequency, as above), while controlling for children's age in months (scaled and centered) and the potential interaction between these two factors:

glmer(Response ~ Condition * Age + (1+Condition|Subject) + (1+Age|TargetSentence), family="binomial", data=Data)

Responses were coded as (1) uninversion error (e.g., **Who he can draw?*; *N*=159) or (0) correct question (e.g., *Who can he draw?*; *N*=647), with all other responses excluded as missing data (*N*=266); hence the use of a binomial outcome variable (logit function). Although the rate of missing data might seem relatively high, it reflects the fact that – due to our focus on particular n-grams – it was necessary to exclude otherwise-scorable questions where children made perfectly reasonable substitutions (e.g., *Who can the man draw?*). Similar numbers of scorable responses were produced in the high-frequency (*N*=415) and low frequency conditions (*N*=391).

The model set out above failed to converge. Thus, in accordance with our pre-registered analysis plan, we removed the by-TargetSentence random slope of Age, which allowed the model to converge. This model is summarized in Table 2 (see Appendix 1 for the full model). A main effect of Age was observed, reflecting the fact that the rate of uninversion errors decreased with development. However, our pre-registered prediction of a main effect of condition (at $p<0.05$) was not supported; neither was a significant interaction of Condition by Age observed[4]. Indeed, children produced uninversion errors at very similar rates in the high-frequency condition (*M*=0.21, *CI*=0.17-0.25) and the low-frequency condition (*M*=0.19, *CI*=0.15-0.22). Note that the study was powered for a small effect size (*d*=0.2), and so we have reason to consider that this is a genuine null effect rather than a false negative.

Table 2. ***Mixed-effects model for the main, pre-registered analysis. Model summary statistics: AIC=584.6, BIC=622.2, logLik=-284.3, deviance=568.6, df.resid=798.***

Fixed Effect	Estimate	*SE*	*z* value	Pr(>|*z*|)
(Intercept)	-2.98	0.62	-4.78	1.72E-06
ConditionLow	0.24	0.65	0.37	0.7083
Age	-0.85	0.42	-2.01	0.0441
Condition-Low:Age	0.31	0.28	1.12	0.2622

TargetSentencePair failed to converge, apparently because the two are so highly correlated. A model that included a random slope for TargetSentencePair but not TargetSentence yielded similar *p* values to the model reported above, for both Condition ($p=0.44$) and Age ($p=0.04$).

[4] The study pre-registration stated that "P-values will be computed via Kenward-Roger and Satterthwaite approximations". However, this method is in fact applicable for continuous dependent variables only. Thus, we instead report *p* values approximated from the *Z* distribution. We also ran a version of the model with no interaction, in order to allow us to compute *p* values for the main effects of Condition and Age via likelihood ratio test (drop1 function of lme4): $p=0.88$ and $p=0.09$ respectively.

Random effects:

Groups	Name	Variance	*SD*	Corr
Subject	(Intercept)	7.9127	2.813	
	ConditionLow	0.2661	0.5158	-1
TargetSentence	(Intercept)	0.8352	0.9139	

It is important to acknowledge at this point that while this null finding was not predicted by the exemplar-focussed variety of input-based account that we set out to test (e.g., Langacker, 1988; Abbot-Smith & Tomasello, 2006; Goldberg, 2006; Ambridge 2020a, 2020b; McCauley et al., 2021) it is potentially consistent with slot-and-frame-focussed input-based accounts which would seem to assume "free slots" in the crucial Bigram 3 position (e.g., *What+can **[THING] [PROCESS]**?*) (e.g., Rowland & Pine, 2000; Rowland, Pine, Lieven & Theakston, 2003, 2005; Dabrowka & Lieven, 2005; Ambridge, Rowland, Theakston, & Tomasello, 2006; Rowland, 2007; Ambridge & Rowland, 2009). On the other hand, this strict reading of slot-and-frame accounts is difficult to reconcile with the findings of McCauley et al. (2021) of frequency effects in the second, third and fourth bigram positions (spanning the "free slot").

The source of this discrepancy is not easy to pinpoint, but one possibility is that the present dataset does, in principle, include evidence for frequency effects spanning the "free slot', just not necessarily in the third bigram position. We explore this possibility in a series of non-preregistered, exploratory analyses.

Exploratory analyses

The main, pre-registered analysis reported above failed to find any evidence of an effect of the frequency of the third bigram from uninverted structures (e.g., **Who he **can draw/name**?*) on rates of uninversion error. In this analysis, as preregistered, the frequency of the third bigram was treated as a categorical predictor and other n-grams kept constant across the paired items as much as possible. However, there is variance between items beyond these pairs and given that several previous studies have found frequency effects in multiple positions (for n-grams from both inverted and uninverted structures), we conducted a series of exploratory analyses designed to investigate whether any of these effects are observed in the present dataset. Although researcher degrees of freedom are always a concern in non-preregistered analyses (Simmons, Nelson, & Simonsohn, 2011), these are minimized by the fact that our analysis strategy is identical to that of McCauley et al., (2021), with all analyses conducted on the main dataset from the preregistered analysis, with no further exclusions, transformations, recodings etc.

There are various challenges in these analyses, given that the stimuli were not designed to look at these effects, but rather effects within high-/low-frequency matched pairs. Many n-gram frequencies were correlated with one another, creating a problem of multicollinearity. Furthermore, since the present stimuli include just 16 questions (and just 8 *wh*-word+subject+auxiliary combinations), we have a very low ratio of items to predictor variables, which also makes it more difficult to statistically tease apart these predictors (cf., McCauley et al., 2021). Thus, these analyses should be

treated as highly exploratory, and will require confirmation from future suitably designed studies.

In order to address these difficulties, we first took the decision to disregard trigrams, and investigate only the question of whether bigram effects are observed above and beyond unigram (single-word frequency) effects. Excluding trigrams reduces both the problem of collinearity (since trigram frequency is correlated with the frequency of its component unigrams and bigrams) and the low item:predictor ratio (by removing predictors).

We first fit a full model with all unigrams and bigrams (inverted and uninverted) as fixed effects, random effects of participant on the intercept and all slopes, and a random effect of sentence on the intercept. This model did not converge and so we simplified by removing the correlation between the participant random effects. We also excluded uninverted bigram 2 as lme4 determined it to be causing rank-deficiency, presumably because of multicollinearity. This model converged although many of the random effects were returned as zero due to their very small size. To give greater stability throughout our inference process we removed all random effects for slopes that were returned as zero. The random effect of sentence on the intercept was also returned as zero but we retained it in the model in order to be maximally conservative in testing for effects.

In order to see whether any of the n-grams had unique explanatory value with regards to the children's errors, we performed a drop-one analysis where we took the all-predictor model and dropped each n-gram fixed effect in turn, looking at whether doing so hurt fit using a likelihood ratio test. If so then we concluded that it was accounting for unique variance in the full model. The final model is shown in Table 3, which also shows the *p* values from the likelihood ratio (drop1) test. The fixed effects (log_) B1, B2 and B3 refer respectively to the (log-transformed) frequency of the first, second and third bigrams from inverted questions; B1.U, and B3.U of the first and third bigrams from uninverted questions (recall that the second was already excluded earlier). Fixed effects of the (log) frequency of individual words (i.e., unigrams U1, U2, U3 and U4) were included in order to allow us to test whether the frequency of a given individual bigram *combination* explained variance above and beyond the frequency of the individual words that make up that bigram.

This analysis tells us (using the example target question *What are you doing?*) that unigrams 1 and 2 (e.g., *what, are*), inverted bigrams 2 and 3 (*are+you, you+doing*) and uninverted bigram 1 (*what+you*) explain unique variance, with the likelihood of a non-inversion error (the dependent measure) decreasing as a function of the inverted bigram frequency (log_B2, log_B3) and increasing as a function of uninverted bigram frequency (log_B1.U).

Checking for a unique effect of the n-grams is an appropriately conservative way of proceeding. However, is it important to note that, due to collinearity, the absence of a unique effect for any given n-gram could simply be the result of its not being separable from other variables in this particular dataset. In order to look at the theoretical separability of the predictors, we performed Principal Components Analysis (PCA). PCA is a dimensionality-reduction algorithm that, when given a matrix of variables –

Table 3. ***Bigram predictors in exploratory analysis all n-gram model. P values are based on the chi-square (likelihood ratio test) drop-one method. (log_) B1, B2 and B3 refer respectively to the (log-transformed) frequency of the first, second and third bigrams from inverted questions; B1.U, B2.U and B3.U of the first, second and third bigrams from uninverted questions; U1, U2, U3 and U4 of unigrams (i.e., single words). Model summary statistics: AIC=522.9, BIC=593.3, logLik=-246.5, deviance=492.9, df.resid=791.***

Fixed Effect	**M**	**SE**	***p*_drop1**
(Intercept)	-4.1291	0.7314	NA
log_U1	**12.0778**	**3.3484**	**0.0001246**
log_U2	**1.9601**	**1.0172**	**0.04982**
log_U3	-1.7205	1.1222	0.1222
log_U4	1.3875	1.0038	0.1617
log_B1	-0.8387	0.7302	0.2494
log_B2	**-1.883**	**0.5308**	**0.0001334**
log_B3	**-1.927**	**0.8685**	**0.0229**
log_B1.U	**15.2218**	**4.4347**	**0.0002763**
log_B3.U	0.1845	0.1761	0.2925

Random effects:

Groups	**Name**	**Variance**	**Std.Dev.**
Subject	(Intercept)	1.20E+01	3.47E+00
Subject.1	log_U1	4.32E-14	2.08E-07
Subject.2	log_U3	3.27E+00	1.81E+00
Subject.3	log_B1.U	3.39E-01	5.82E-01
TargetSentence	(Intercept)	2.88E-15	5.37E-08

in this case, n-gram predictor variables – collapses highly correlated variables into composite variables ("components"). By looking at how the original variables load onto these components, we can observe how separable they are. Figure 3 shows the loading of all variables onto the first two components, which account for 47% and 27% of the variance respectively. B1, B2 and B3 refer to the first, second and third bigrams from inverted questions; B1.U, B2.U and B3.U to the first, second and third bigrams from uninverted questions; U1, U2, U3 and U4 to unigrams. The further away any two variables are, particularly along the horizontal axis (the first compo accounts for more of the shared variance across the predictors), the more separable they are. It is clear that B1 and B2, being very close together, are hard to separate. It is plausible then that B1 does explain variance in the children's production, but this was a subset of the variance explained by B2, and thus we saw no unique effect of B1. The same applies for B1.U and B2.U, which could explain why B2.U was rejected as rank deficient. A similar situation can be seen for U1 and U3, which could explain why U3 was not found to explain unique variance, and U4 and B3, which could explain why only

the latter explained unique variance. Finally, the very close proximity between U2 and B3.U could explain why only the former is seen to explain unique variance.

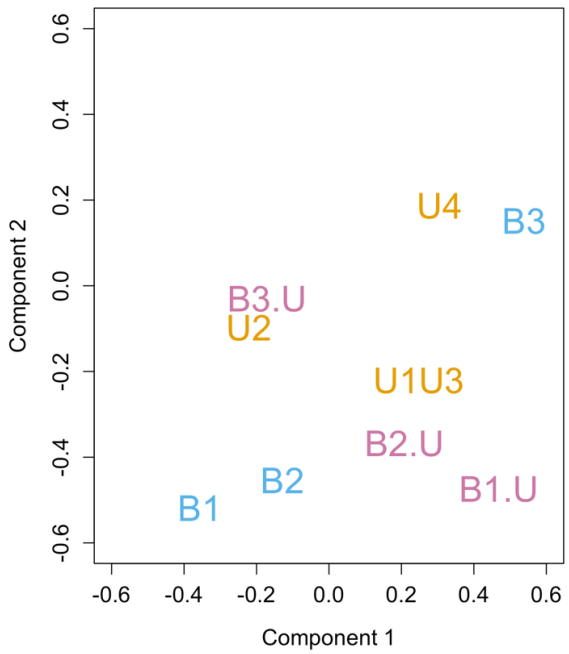


Figure 3: ***Loading of N-gram frequency variables on the first two principal components, which account for 47% and 27% of variance respectively. Unigrams appear in orange (U1, U2, U3, U4), inverted bigrams in light blue (B1, B2, B3) and univerted bigrams in pink (B1.U, B2.U, B3.U). B1, B2 and B3 refer respectively to the first, second and third bigrams from inverted questions; B1.U, B2.U and B3.U to the first, second and third bigrams from uninverted questions; U1, U2, U3 and U4 to unigrams (i.e., single words). The further away any two variables, particularly along the horizontal axis, the greater the extent to which they are separable.***

Summary of Exploratory Effects.

Consistent with a frequent claim in the literature (e.g., Rowland & Pine, 2000; Rowland, 2007; Ambridge & Rowland, 2009), the present exploratory analysis found preliminary evidence that children make fewer uninversion errors for questions that contain bigrams with high input frequency. Although preliminary, this evidence is important, as it is the first experimental study (cf. the corpus study of McCauley et al., 2021) to demonstrate the existence of bigram effects in question production *above-and-beyond effects of the component unigrams*. That is, children make fewer uninversion errors when the bigrams that make up the question (e.g., *can he…*; *…he draw…*) are of higher frequency, independent of the frequency of the individual words (*can, he, draw*). Echoing the corpus study of McCauley et al. (2021), we also found evidence that rates of uninversion error (e.g., **Who he can draw?*) are higher when the uninverted bigrams (e.g., *Who he…*) are of higher frequency in the input. It is important to treat the ***specific*** effects seen with some caution and note that while we saw unique effects of some predictors and not others, the absence of an effect could in part be the effect of collinearity — a variable can spuriously appear not to have an effect

because its variance is being explained by another variable with which it is collinear — and it could be that in a set of stimuli where the variables were more separable we would see different specific patterns. For this reason, we have avoided the temptation of speculating as to potential reasons why it might be *these* particular bigrams that seem to yield frequency effects and not others. Importantly, however, the conclusion that *some* n-gram frequencies are predictive of errors rates is not affected by collinearity, which affects only our ability to say *which ones*.

Discussion

Consistent with input-based accounts of question acquisition, several previous studies have shown that children are less likely to produce uninversion errors (e.g., **Who he can draw?*) when lexical strings that appear in the correct form – particular *wh*-word+auxiliary combinations such as *who can* – are frequent in the input (e.g., Rowland & Pine, 2000; Rowland, Pine, Lieven, & Theakston, 2003, 2005; Dabrowka & Lieven, 2005; Ambridge, Rowland, Theakston, & Tomasello, 2006; Rowland, 2007; Ambridge & Rowland, 2009). Ambridge and Rowland (2009) and McCauley, Bannard, Theakston, Davis, Cameron-Faulkner, and Ambridge (2021) also showed that children are more likely to produce uninversion errors (e.g, **Who he can draw*) when lexical strings that appear in the errorful form are frequent in the input (e.g., *he can* is considerably more frequent than *can he*).

The aim of the present study was to conduct a preregistered experimental test of a prediction that follows from the study of McCauley et al. (2021), and from the more general claim of input-based approaches that learners retain, and are influenced by, individual lexical strings even when they have formed more abstract representations too: Participants will produce more uninversion errors when those errors incorporate – in the Bigram 3 position – high-frequency bigrams from uninverted structures (e.g., **Who he* ***can draw****?*) than lower-frequency bigrams from uninverted structures (e.g., **Who he* ***can name?***); with all other bigrams and unigrams (i.e., single words) either identical (e.g., *Who+he, he+can, he, can*) or closely matched for frequency (e.g., *draw* and *name*). The present study tested this prediction using an elicited-production paradigm in which children put questions to a talking dog toy.

This main, preregistered prediction was not supported. Given that the study was well powered (67 participants, yielding 90% power *a priori*) for even a small effect size (d=0.2), these findings are plausibly consistent with a genuine null effect rather than a false negative. Given that this effect has been seen in naturalistic data (McCauley et al., 2021), it is somewhat surprising that we failed to find it in this experiment. One possible explanation for this discrepancy is that in order to control for the frequency of other component n-grams, we were forced to select items in which there was inadequate difference between the frequency of bigram 3 in the inverted and the uninverted form. On this view, McCauley et al.'s (2021) finding of a frequency effect in (amongst others) the third bigram position reflects a genuine effect, and the present null finding is a result of methodological factors. Of course, it is also possible that the opposite is true: Whenever an effect is found in observational data but not replicated in an experiment, the possibility exists that the apparent effect in the former is due to unmeasured confounding. A third possibility, and the one we favour, is that whether or not frequency effects are observed for a given n-gram position depends

on factors such as the particular lexical question forms under investigation, and participant-level factors such as linguistic history, memory and willingness to generalize beyond the input. In all likelihood, the only way to resolve this issue will be to build detailed computational models that make specific predictions regarding specific lexical question types (possibly for specific individuals), rather than naïve n-gram models that predict equivalent frequency effects across the board.

We follow our pre-registered analysis with non-preregistered exploratory analyses of the data. In this we explored frequency effects for other n-grams. It is important to note that the stimuli were not designed to look at these effects and thus they are confounded with covariance in other n-grams. Nevertheless, we found evidence of a facilitatory effect on correct-question production of the frequency of the second and third bigrams from inverted structures (e.g., *can he…he draw*), even after controlling for unigram frequency (unlike, for example, Ambridge & Rowland, 2009). The frequencies of the first and second bigrams were highly correlated so that it is possible that an effect of the first bigram was hidden. We also saw evidence that rates of uninversion error (e.g., **Who he can draw?*) are higher when the first uninverted bigram (e.g., *Who he…*) are of higher frequency in the input.

Before moving on to explore the potential theoretical implications of the present findings, it is important to acknowledge three possible methodological objections. The first is that – as a result of the tight constraints imposed by the need to match stimuli in the high- and low-frequency conditions – some of the target questions were rather unnatural and/or difficult to illustrate with pictures. It is true that some of the questions are somewhat unnatural, although we did our best to ameliorate any unnaturalness as far as possible with the preliminary lead-in sentences (e.g., *In this jigsaw, it looks like he means something. I wonder what he can mean*). Interested readers are invited to draw their own conclusions regarding the extent to which we succeeded by perusing our full list of prompts, which can be found in Appendix 3 (pictures can be found on the accompanying OSF site at https://osf.io/74urw/. We do not consider it appropriate, however, to conduct an item analysis since our target questions vary with regard to properties other than their perceived naturalness – most importantly the n-gram statistics used as fixed-effect predictors in our exploratory analyses – and one advantage of mixed-effects models is that they allow us to *control for* item-by-item differences that are not captured by the fixed-effects (including naturalness, the particular subject used in the question, differences relating to the illustrations etc.).

The second potential methodological objection is that (as already mentioned in the Methods section), by including uninverted question strings in the experimenter's prompt ("Let's ask the dog *where Daddy is sitting*. You ask the dog *where Daddy is sitting*") we primed children to produce uninverted questions (e.g., **Where Daddy is sitting?*). It is almost certainly the case that such priming will have occurred, since both abstract and lexical priming effects are well established for young children (e.g., Savage et al., 2003; Huttenlocher et al., 2004; Bencini & Valian, 2008; Rowland et al., 2012). The question is whether this priming effect replaced and supplanted children's normal mechanisms of question production to the extent that the present (tentative) findings of certain n-gram effects are entirely invalidated. We do not believe this to be the case for three reasons. First, overall, children produced around four times as many correct as uninverted questions. Clearly, then, children's normal production

mechanisms were, on the whole, operating well; indeed, four times out five, they were able to override any priming effect. Second, although this rate of uninversion errors (20%) is much higher than rates observed in naturalistic data (e.g., McCauley et al., 2021, found just 2%), this is not a fair comparison, since naturalistically-produced questions follow a broadly Zipfian distribution with just a handful of potentially-rote questions (e.g., *What's that?; What are you doing?*) accounting for the majority of all tokens. When we control for this skewed distribution by counting types not tokens, uninversion errors also occur at a rate of around 20% in naturalistic data (e.g., Rowland & Pine, 2000), suggesting that the present method does not artificially inflate rates of universion error; or at least, not to a great extent. Third, at a broad-brush level, the findings of certain n-gram effects on rates of uninversion error echo those of McCauley et al. (2021), which were based entirely on corpus data. Overall, then, we feel justified in claiming that while the experimenter's prompt certainly encouraged children to produce uninversion errors – to some extent, that was exactly the aim – it did so in a way that elucidates, rather than obscures, underlying question-by-question differences in rates of uninversion error versus correct questions.

The third potential methodological objection that we must consider is that by excluding questions that did not use the precise target words (e.g., if the child said, "Who can the man draw?" rather than "Who can he draw?"), we incorrectly estimated overall rates of uninversion errors versus correct questions This is true, but it was never our intention to make any theoretical claims on the basis of *overall* rates of uninversion errors versus correct questions, and, indeed, we do not do so. Any such claim would be problematic given the finding from both the present study (tentatively) and previous studies (more securely) that error rates vary dramatically by question type (e.g., Rowland & Pine, 2000, report uninversion rates of 100% for some questions and 0% for others). Thus, the overall rate of uninversion errors versus correct questions in any particular experimental study is determined, at least to a considerable degree, by the particular question types chosen, meaning that any theoretical claims based on *overall* error rates would invariably be misleading. Relatedly, we do not see it as a problem that particular *wh*-words and particular auxiliaries appeared at unequal rates in our stimuli (which was necessary in order to create closely-matched high-/low-frequency pairs); since at no point do we analyse – much less make claims on the basis of – error rates at the by-*wh*-word or *by*-auxiliary level.

Returning now to the present findings and their implications, when taken together with the findings of McCauley et al. (2021), the exploratory findings from the present study suggest that children's language production mechanism is sensitive to unigram, bigram and trigram frequency, even when those strings are from very different sentence types to the target. That is, strings from declarative input utterances affect the production of questions; specifically, by increasing rates of uninversion error. What is less clear is whether material at the left-hand edge of questions is somehow privileged (e.g., *What are you…*) or, conversely, whether n-grams further to the right from both inverted (e.g., *you doing; doing there*) and uninverted structures (e.g., *you are, are doing, doing there*) play a large – or even larger – role.

Certainly, neither the present findings nor those of McCauley et al. (2021) are consistent with a strict interpretation of proposals such as Rowland and Pine (2000), Dabrowska and Lieven (2005), Ambridge et al. (2006) and Ambridge & Rowland (2009)

that children's early question schemas are of the form *What are [THING] [PROCESS]?* That is, the findings of the present study and McCauley et al. (2021) are not consistent with a "left-edge bias" view under which only the *wh*-word+auxiliary combination is frozen as a learned schema, with the [THING] and [PROCESS] slots entirely "free" (a *strict* interpretation of these previous proposals; and not necessarily an interpretation that their authors would endorse).

In fact, some of the previous evidence for a special role for *wh*-word+auxiliary combinations may not be as strong as it first appears. For example, while Ambridge & Rowland's (2009) experimental study found a negative correlation between children's rates of uninversion error and the input frequency of *wh*-word+auxiliary combinations (or, for *yes/no* questions auxiliary+subject combinations) this correlation held only when removing the outlier Why+can which shows much lower rates of error than would be predicted by its very low input frequency. The corpus study of Rowland and Pine (2000) did not in fact test for this correlation at all, but instead provided evidence only for the weaker claim that "the wh + aux combinations the child uses are more frequent in the mother's input than those the child fails to use (i.e. that occur divided by a subject in uninverted *wh*-questions)". Westergaard (2009) further argues that (1) Many of the child's uninverted questions should have been excluded from Rowland and Pine's (2000) analysis because they were produced only once or a handful of times and (2) Many of the child's inverted questions should have been excluded, because they include the dummy auxiliary DO (e.g., *What does; Where did*) which children already know – for quite independent reasons – is not normally included after a subject unless for emphasis. For example, a child would not normally say *He does like it* or *He did go to school* (cf., *He likes it; He went to school)* rendering the non-occurrence of *What he does like?* or *Where he did go?* moot.

The most convincing evidence for a special role for the left-edge of the utterance comes from the corpus study of Rowland (2007, which found a significant negative correlation between the frequency of the frame (again defined as wh+auxiliary for wh-questions and auxiliary+subject for yes/no questions) and rates of uninversion errors (versus correct questions), over and above auxiliary type (DO vs modal). However, this study did not control for the independent input frequency of other bigrams in the well-formed question, or of unigrams.

Recall, too, that the present study does not constitute strong evidence against a special role for the first bigram (here, always *wh*-word+auxiliary) due to collinearity between the input frequency of the first bigram (e.g., *what+are*), which was not a significant predictor of correct production, and the second bigram (*are+you*), which was. Thus, the jury is still very much out with regard to the question of whether the n-grams at the left edge of the utterance hold some special importance for question acquisition (e.g., by leading to the formation of slot-and-frames patterns like *What are [THING] [PROCESS]*)? It is also important to emphasize that while Rowland and Pine (2000) and Rowland (2007 focussed on the *protective* effect of high-frequency inverted strings on correct question production, the present study (like Ambridge & Rowland, 2009; McCauley et al., 2021) additionally investigated the potentially error-causing (lure) effect of high-frequency uninverted strings on uninversion errors.

What the present exploratory findings tentatively suggest is that at a general level (i.e., setting aside the question of a left-edge bias), n-gram frequency indeed affects the relative probability of uninversion errors versus correct-question production: high-frequency n-grams with inverted order pull towards correct questions; high-frequency n-grams with uninverted word order pull towards non-inversion errors. Thus the present findings – like those of McCauley et al. (2021) – are consistent with a view under which, having generalized in some sense across input utterances to yield more abstract representations, traces left by the initial input utterances are not discarded but retained (in principle, forever), and influence subsequent language production and processing (e.g., Langacker, 1988; Abbot-Smith & Tomasello, 2006; Goldberg, 2006; Ambridge 2020a, 2020b). As discussed in Ambridge (2020b) the generalizations yielded by such a model are likely to exist at numerous levels of abstraction simultaneously, and may look very different to the types of generalizations posited under traditional linguistic analysis (e.g., a *[WH-WORD] [AUXILIARY] [THING] [PROCESS]*? construction or a *subject-auxiliary inversion* rule). Indeed, it is important to emphasize that at a global level (we are not aware of any studies looking specifically at adult question production) frequency effects are ubiquitous not just in child language acquisition (e.g., Ambridge et al., 2015), but in adult language processing too (in addition to the studies cited in the Introduction, see e.g., the summaries by Ellis, 2002; Gries & Divjak, 2012). Frequency effects – including n-gram effects – are not solely a hallmark of child language acquisition that disappear when more abstract representations are formed. Rather, what we need are accounts that can explain both abstract and lexical effects at once, for both adults and children.

On this note, it is important to reiterate, as stated in the Introduction, that the present findings (and McCauley et al., 2021) do not demonstrate the *absence* of a syntactic subject-auxiliary inversion rule. What they do suggest is that proponents of such accounts owe an explanation as to the source of the observed unigram, bigram, and trigram effects; for example, in terms of the filtering of a *subject-auxiliary inversion* rule through a production mechanism that is sensitive to n-gram frequency. Note that this is only a suggestion; we are not aware of any rule-based accounts of the acquisition of question production that actually incorporate such a mechanism.

In turn, researchers who advocate the abandoning of accounts based on the notion of a subject-auxiliary inversion rule owe an account of exactly how children acquire the ability to move beyond the n-gram strings that they hear in the input and develop abstract representations that allow them to produce entirely novel questions (including those for which many individual n-gram frequencies will be zero, or at least extremely low).

At present, descriptive verbal accounts – on both the rule-based and construction-based sides – do not make sufficiently precise quantitative predictions that they can be subjected to objective empirical testing. For example, as we have noted throughout, it is not clear whether slot-and-frame-based accounts really predict the absence of frequency effects in "free slot" position (e.g., *What+can [THING] [PROCESS]?*), or even – necessarily – their attenuation. If precise quantitative predictions are to be derived from accounts of question acquisition, then it will almost certainly be necessary to implement these accounts as mechanistic computational models.

In the meantime, while the present study – contra McCauley et al. (2021) – found no evidence for the special importance in question formation of the third bigram from uninverted utterances, it does suggest that children's question production is indeed influenced by unigram, bigram, and trigram frequency; findings that any successful account of children's question acquisition – and of their language acquisition more generally – will need to explain.

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*(3), 275-290.

Ambridge, B. (2020a). Against stored abstractions: A radical exemplar model of language acquisition. *First Language, 40*(5-6), 509-559.

Ambridge, B. (2020b). Abstractions made of exemplars or 'You're all right and I've changed my mind' Response to commentators. *First Language, 40*(5-6), 640-659.

Ambridge, B., & Rowland, C.F. (2009). Predicting children's errors with negative questions: Testing a schema-combination account. Cognitive Linguistics, 20(2), 225-266.

Ambridge, B., Rowland, C. F., & Pine, J. M. (2008). Is Structure Dependence an innate constraint? New experimental evidence from children's complex question production. Cognitive Science, 32(1), 222-255.

Ambridge, B., Rowland, C. F., Theakston, A. & Tomasello, M. (2006) Comparing Different Accounts of Non-Inversion Errors in Children's Non-Subject Wh-Questions: 'What experimental data can tell us?' *Journal of Child Language* 30(3) 519-557.

Ambridge, B., Rowland, C.F., Theakston, A.L. & Kidd, E.J. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language, 42*(2), 239-73.

Arnon, I. & Clark, E. V. (2011). When 'on your feet' is better than 'feet': Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development, 7*, 107-129.

Arnon, I. & Snider, N. (2010) More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*: 67–82.

Arnon, I., McCauley, S.(C) & Christiansen, M. H. (2017). Digging up the building blocks of language: Age-of-Acquisition effects for multiword phrases. Journal of Memory and Language, 92, 265-280.

Bannard, C. (2006). *Acquiring phrasal lexicons from corpora* (Doctoral dissertation, University of Edinburgh).

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

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: 10.18637/jss.v067.i01.

Bellgui, U. (1971). Simplification in children's language. In R. Huxley & E. In-gram (Eds.), *Language acquisition: Models and methods*. New York: Academic Press.

Bencini, G. M. L., & Valian, V. V. (2008). Abstract sentence representations in 3 year-olds: Evidence from language production and comprehension. *Journal of Memory and Language, 59*, 97 - 113.

Bloom, L., Merkin, S., & Wooten, J. (1982). Wh-Questions: Linguistic factors that contribute to the sequence of acquisition. *Child Development, 53*, 1084-1092.

Bloom, L., Merkin, S., & Wooten, J. (1982). Wh-Questions: Linguistic factors that contribute to the sequence of acquisition. *Child Development, 53*, 1084-1092.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: the reduction of don't in English. *Linguistics, 37*(4), 575-596.

Dabrowska, E., & Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics, 16*(3), 437-474.

DeVilliers, J. (1991). Why question? In T. L. Maxfield & B. Plunkett (Eds.), *Papers in the acquisition of wh: Proceedings of the UMASS Roundtable, May 1990*. Amherst, MA: University of Massachusetts Occasional Papers.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition, 24*(2), 143-188.

Erreich, A. (1984). Learning how to ask: Patterns of inversion in yes-no and wh-questions. *Journal of Child Language, 11*, 597-592.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. New York: Oxford University Press.

Gries, S. T., & Divjak, D. (Eds.). (2012). *Frequency effects in language learning and processing*. De Gruyter Mouton.

Hattori. (2003). Why do children say did you went?: the role of do-support. *Supplement to the Proceedings of the 28th Boston University Conference on Language Development*. (http://www.bu.edu/linguistics/APPLIED/BUCLD/supp.html)

Havron, N., & Arnon, I. (2021). Starting big: The effect of unit size on language learning in children and adults. *Journal of Child Language*, *48*(2), 244-260.

Hernández, M., Costa, A., & Arnon, I. (2016). More than words: multiword frequency effects in non-native speakers. *Language, Cognition and Neuroscience*, *31*, 785-800.

Hurford, J. (1975). A child and the English question formation rule. *Journal of Child Language, 2*, 299-301.

Huttenlocher, J., Vasilyeva, M., & Shimpi, P. (2004). Syntactic priming in young children. *Journal of Memory and Language, 50*(2), 182-195.

Janssen, N. & Barber, H.A. (2012) Phrase frequency effects in language production. PLoS ONE 7(3): e33202. doi:10.1371/journal.pone.0033202.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. in Bybee, Joan and Paul Hopper (eds.). 2000. Frequency and the emergence of linguistic structure. Amsterdam: John Benjamins (pp. 229-254).

Kueser, J.B, & Leonard, L.B. (2020. The Effects of frequency and predictability on repetition in children with Developmental Language Disorder. *Journal of Speech Language and Hearing Research, 63*(4):1165-1180. doi: 10.1044/2019_JSLHR-19-00155.

Krug, M. (1998). String frequency. A cognitive motivating factor in coalescence, language processing, and linguistic change. *Journal of English Linguistics, 26*, 286–320.

Kuczaj, S. (1976). Arguments against Hurford's 'Aux copying rule'. *Journal of Child Language, 3*, 423-427.

Kuczaj, S. A., & Brannick, N. (1979). Children's use of the wh question modal auxiliary placement rule. *Journal of Experimental Child Psychology, 28*, 43-67.

Labov, W., & Labov, T. (1978). Learning the syntax of questions. In R. Campbell & P. Smith (Eds.), *Recent advances in the psychology of language*. New York: Plenum Press.

Langacker, R.W., (1988). A usage-based model. In B. Rudzka-Ostyn (ed.), Topics in Cognitive Linguistics. Amsterdam: John Benjamins, pp. 127-161.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech, 6*(3), 172-187.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database.* Mahwah, NJ: Lawrence Erlbaum Associates.

Maratsos, M., & Kuczaj, S. (1978). Against the transformationalist account: A simpler analysis of auxiliary overmarking. *Journal of Child Language, 5*, 337-345.

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

McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities. *Psychological Science, 14*, 648–652.

Pluymaekers, M., Ernestus, M., & Baayen, R. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica, 62*(2-4), 146-159.

Pozzan, L., & Valian, V. (2017). Asking questions in child English: Evidence for early abstract representations. *Language Acquisition, 24*(3), 209-233.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ringstad, T. & Kush, D. (2021) Learning embedded verb placement in Norwegian: Evidence for early overgeneralization. *Language Acquisition.*

Rowland, C. F. (2007). Explaining errors in children's questions. *Cognition*, *104*(1), 106-134.

Rowland, C. F., & Pine, J. M. (2000). Subject-auxiliary inversion errors and wh-question acquisition: 'What children do know?' *Journal of Child Language, 27*(1), 157-181.

Rowland, C. F., Pine, J. M., Lieven, E. V. M., & Theakston, A. L. (2003). Determinants of acquisition order in wh questions: re-evaluating the role of caregiver speech. *Journal of Child Language,* 609-635.

Rowland, C.F., Pine, J. M., Lieven, E. V. M., & Theakston, A. L. (2005). The incidence of error in young children's wh-questions. *Journal of Speech, Language and Hearing Research, 48* 384-404.

Rowland, C.F., Chang, F., Ambridge, B., Pine, J. M., & Lieven, E.V.M. (2012). The development of abstract syntax: Evidence from structural priming and the lexical boost. *Cognition, 125*(1) 49–63.

Santelmann, L., Berk, S., Austin, J., Somashekar, S., & Lust, B. (2002). Continuity and development in the acquisition of inversion in yes/no questions: dissociating movement and inflection. *Journal of Child language, 29*(4), 813-842.

Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and Language, 162,* 42-45.

Savage, C., Lieven, E., Theakston, A., & Tomasello, M. (2003). Testing the abstractness of children's linguistic representations: lexical and structural priming of syntactic constructions in young children. *Developmental Science, 6*(5), 557-567.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22*(11), 1359-1366.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W.J.B. (2011) Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition 37*: 776.

Skarabela, B., Ota, M., O'Connor, R., & Arnon, I. (2021). 'Clap your hands' or 'take your hands'? One-year-olds distinguish between frequent and infrequent multiword phrases. *Cognition, 211,* 104612.

Sosa, A.V. & MacFarlane, J. (2002) Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word of. *Brain and Language 83*: 227–236.

Stromswold, K. (1990). *Learnability and the acquisition of auxiliaries.* Unpublished Ph.D. dissertation, MIT.

Stromswold, K. (1995). The acquisition of subject and non-subject wh-questions. *Language Acquisition, 4*(1), 5-48.

Tremblay, A. and Baayen, R. H. (2010) Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood, *Perspectives on formulaic language: Acquisition and communication*. London: The Continuum International Publishing Group

Tyack, D., & Ingram, D. (1976). Children's production and comprehension of questions. *Journal of Child Language, 4*, 211-224.

Valian, V., & Casey, L. (2003). Young chilren's acquisition of wh-questions: the role of structured input. *Journal of Child Language, 30,* 117-143.

Valian, V., Lasser, I., & Mandelbaum, D. (1992). *Children's early questions.* Paper presented at the 17th Annual Boston University Conference on Language Development, Boston, MA.

Waldmann, C. (2011). Moving in small steps towards verb second: A case study. *Nordic Journal of Linguistics, 34*(3), 331-359.

Westergaard, M. (2009). Usage-based vs. rule-based learning: the acquisition of word order in wh-questions in English and Norwegian. *Journal of Child Language, 36*(5), 1023-1051.

Westergaard, M. & K. Bentzen. (2007). The (non-) effect of input frequency on the acquisition of word order in Norwegian embedded. *Frequency effects in language acquisition: Defining the limits of frequency as an explanatory concept*, *32*, 271.

Data, code and materials availability statement

All raw data, analysis code (for R) and materials (a package for the Open Source Python package, Open Sesame: https://osdoc.cogsci.nl) can be downloaded from https://osf.io/74urw/

Ethics statement

Ethics approval was obtained from the University of Liverpool Research Ethics Committee prior to recruitment. Children's caregivers gave informed written consent and children gave verbal assent.

Authorship and Contributorship Statement

BA, SM, CB, TC-F and AT conceived of the study and designed the study. SM and BA wrote the first draft of the manuscript. AG contributed to the design of the study and collected the data. SM, CB and BA analyzed the data. BA, SM, CB and AT revised the manuscript. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

This work was supported by the International Centre for Language and Communicative Development (LuCiD). The support of the Economic and Social Research Council [ES/L008955/1 and ES/S007113/1] is gratefully acknowledged.

Appendix 1: Full R output for the main preregistered analysis

```
[1] "Now the preregistered model: We said 'In the event of convergence
    failure, we will simplify the model by simplifying the random effects
    terms to no longer include the by-subject random slope for condition
    or the by-item random slope for age. In the event of further conver-
    gence failure we will remove the fixed effect of subject age'"
[1] "Here's a summary of the final model - We had to drop the by-Tar-
    getSentence random slope for Age"
Generalized linear mixed model fit by maximum likelihood (Laplace Approxi-
    mation) ['glmerMod']
 Family: binomial  ( logit )
Formula: Response ~ Condition * Age + (1 + Condition | Subject) + (1 |
    TargetSentence)
   Data: Data

     AIC      BIC   logLik deviance df.resid
   584.6    622.2   -284.3    568.6      798

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9240 -0.2798 -0.1391 -0.0620  4.9796

Random effects:
 Groups         Name         Variance Std.Dev. Corr
 Subject        (Intercept)  7.9127   2.8130
                ConditionLow 0.2661   0.5158   -1.00
 TargetSentence (Intercept)  0.8352   0.9139
Number of obs: 806, groups:  Subject, 67; TargetSentence, 16

Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.9802     0.6229  -4.784 1.72e-06 ***
ConditionLow       0.2448     0.6543   0.374   0.7083
Age               -0.8488     0.4217  -2.013   0.0441 *
ConditionLow:Age   0.3144     0.2804   1.121   0.2622
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
optimizer (Nelder_Mead) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')

[1] "In the preregistration, we said 'P-values will be computed via Ken-
    ward-Roger and Satterthwaite approximations', but this isn't actually
    possible for binomial models So we'll just report the p values from
    the main model output (approximated via the z distribution"
[1] "As a double-check, we'll remove the interaction, which will allow us
    to get p values via drop1, and report this in a footnote"
Single term deletions

Model:
Response ~ Condition + Age + (1 + Condition | Subject) + (1 |
    TargetSentence)
          npar    AIC     LRT Pr(Chi)
<none>         583.93
Condition    1 581.95 0.02461 0.87534
Age          1 584.82 2.89179 0.08903 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
[1] "Now, the pre-registered syntax is all very well, but it seems to me
    (Ben) that we should also include pair ('Set') as a random effect,
    since the high/low frequency manipulation is indeed within each pair,
    and again report it in a footnote"
[1] "Just fails as they're too correlated"
[1] "Probably makes more sense than the pre-registered syntax, but doesn't
    actually change the result at all"
Generalized linear mixed model fit by maximum likelihood (Laplace Approxi-
    mation) ['glmerMod']
 Family: binomial  ( logit )
Formula: Response ~ Condition * Age + (1 + Condition | Subject) + (1 |
    Set)
   Data: Data

     AIC      BIC   logLik deviance df.resid
   572.3    609.8   -278.2    556.3      798

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9853 -0.2855 -0.1379 -0.0594  5.2268

Random effects:
 Groups  Name         Variance Std.Dev. Corr
 Subject (Intercept)  8.3877   2.8962
         ConditionLow 0.3166   0.5627   -1.00
 Set     (Intercept)  0.9492   0.9743
Number of obs: 806, groups:  Subject, 67; Set, 8

Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.0690     0.6402  -4.794 1.64e-06 ***
ConditionLow       0.3380     0.4344   0.778   0.4365
Age               -0.8952     0.4320  -2.072   0.0382 *
ConditionLow:Age   0.3540     0.2775   1.276   0.2020
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 2: Full R output for the exploratory analyses

```
[1] "First model includes the frequency of all unigrams - i.e.,  each of the individual words - called
    log_U1/U2/U3/U4 - all bigrams from the inverted form of the question, called log_B1/B2/B2, and all bigrams from
    the uninverted form of the question, called log_B1.U/B2.U/B3.U. We attempt to include a by-participant random
    slope for all of these predictors, but as we'll see later this won't converge. There are no possible by-Tar-
    getSentence random slopes"
[1] "Doesn't converge so simplify - starting by removing the correlations between the random effect of structure.
    Also remove B2.U as glmer rejects: fixed-effect model matrix is rank deficient"
[1] "Now converges but gives a singular fit. To improve stability for model comparisons, remove all random effects
    that explain close to zero variance (shows up as 0.000e+00) except for TargetSentence which we retain for rea-
    sons of conservatism"
[1] "Still a singular fit, but that's OK!"
[1] "# Now perform a drop one analysis to look at unique contribution of each of the n-grams"
Data: Data
Models:
M_U1: Response ~ log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_U1   14 535.64 601.33 -253.82   507.64
M      15 522.92 593.30 -246.46   492.92 14.721  1  0.0001246 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Unigram1 is retained in the final model"
Data: Data
Models:
M_U2: Response ~ log_U1 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_U2   14 524.77 590.46 -248.38   496.77
M      15 522.92 593.30 -246.46   492.92 3.8475  1    0.04982 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Unigram2 is (narrowly!) retained in the final model"
```

```
Data: Data
Models:
M_U3: Response ~ log_U1 + log_U2 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
    npar    AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_U3   14 523.31 589.0 -247.66   495.31
M      15 522.92 593.3 -246.46   492.92 2.3891  1     0.1222
[1] "Uingram3 is NOT retained in the final model"
Data: Data
Models:
M_U4: Response ~ log_U1 + log_U2 + log_U3 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_U4   14 522.88 588.57 -247.44   494.88
M      15 522.92 593.30 -246.46   492.92 1.9582  1     0.1617
[1] "Uingram4 is NOT retained in the final model"
Data: Data
Models:
M_B1: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_B1   14 522.25 587.94 -247.12   494.25
M      15 522.92 593.30 -246.46   492.92 1.3268  1     0.2494
[1] "Bigram1 from INVERTED forms is NOT retained in the final model"
Data: Data
Models:
M_B2: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
    npar    AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_B2   14 535.51 601.2 -253.76   507.51
```

```
M      15 522.92 593.3 -246.46   492.92 14.593  1  0.0001334 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Bigram2 from INVERTED forms IS retained in the final model"
Data: Data
Models:
M_B3: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
     npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_B3   14 526.10 591.79 -249.05   498.10
M      15 522.92 593.30 -246.46   492.92 5.1758  1     0.0229 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Bigram3 from INVERTED forms IS retained in the final model"
Data: Data
Models:
M_B1.U: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_B1.U   14 534.15 599.84 -253.07   506.15
M        15 522.92 593.30 -246.46   492.92 13.224  1  0.0002763 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Bigram1 from UNinverted forms IS retained in the final model"
Data: Data
Models:
M_B3.U: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_B3.U   14 522.03 587.72 -247.01   494.03
M        15 522.92 593.30 -246.46   492.92 1.1082  1     0.2925
[1] "Bigram3 from UNinverted forms IS NOT retained in the final model"
```

```
[1] "Recall that Bigram2 from UNinverted forms IS NOT retained in the final model as it was already dropped due to
    colinearity"
[1] "Summary: U1, U2, B2, B3 and B1.U explain unique variance"
[1] "Next do a PCA of the bigrams to understand what is going on"
[1] "principal package doesn't do simple PCA. It does PCA plus rotation, so switching to prcomp which is a built-in
    R function"
[1] "Here's the model summary for Table 3"
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 +
    log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 + log_B1.U ||      Subject) + (1 | TargetSentence)
   Data: Data
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   522.9    593.3   -246.5    492.9      791

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.7235 -0.1977 -0.0735 -0.0191 10.8171

Random effects:
 Groups         Name        Variance  Std.Dev.
 Subject        (Intercept) 1.201e+01 3.466e+00
 Subject.1      log_U1      4.322e-14 2.079e-07
 Subject.2      log_U3      3.265e+00 1.807e+00
 Subject.3      log_B1.U    3.385e-01 5.818e-01
 TargetSentence (Intercept) 2.880e-15 5.367e-08
Number of obs: 806, groups:  Subject, 67; TargetSentence, 16

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.1291     0.7314  -5.645 1.65e-08 ***
log_U1      -12.0778     3.3484  -3.607 0.000310 ***
log_U2        1.9601     1.0172   1.927 0.053991 .
log_U3       -1.7205     1.1222  -1.533 0.125230
log_U4        1.3875     1.0038   1.382 0.166924
log_B1       -0.8387     0.7302  -1.149 0.250691
```

```
log_B2       -1.8830     0.5308  -3.548 0.000389 ***
log_B3       -1.9270     0.8685  -2.219 0.026510 *
log_B1.U     15.2218     4.4347   3.432 0.000598 ***
log_B3.U      0.1845     0.1761   1.047 0.294886
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
         (Intr) log_U1 log_U2 log_U3 log_U4 log_B1 log_B2 log_B3 l_B1.U
log_U1    0.272
log_U2   -0.108 -0.296
log_U3    0.112  0.777 -0.096
log_U4   -0.062 -0.289  0.876 -0.065
log_B1    0.064  0.023 -0.700  0.227 -0.516
log_B2    0.235  0.629 -0.187  0.239 -0.040 -0.231
log_B3    0.133  0.575 -0.781  0.357 -0.812  0.533  0.381
log_B1.U -0.259 -0.988  0.361 -0.820  0.323 -0.152 -0.594 -0.632
log_B3.U -0.045 -0.118  0.408  0.015  0.468 -0.265 -0.094 -0.432  0.142
optimizer (bobyqa) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

Appendix 3: Full text of all prompt questions

Oh look…	**In this jigsaw…**	**I wonder/Let's ask the dog/You ask the dog**
… here's the BOY	… he's naming someone	…who he can name
… here's the BOY	… he's drawing someone	…who he can draw
… here's the BOY	… he always needs something	…what he can need
… here's the BOY	… he always eats something	…what he can eat
… here's the BOY	… it looks like he means something	…what he can mean
… here's the BOY	… it looks like he hears something	…what he can hear
… here's DADDY	… he's singing somewhere	…where Daddy is singing
… here's DADDY	… he's sitting somewhere	…where Daddy is sitting
… here's the CAT	… it's causing something	…what it can cause
… here's the CAT	… it's holding something	…what it can hold
… here's the CAT	… it looks like it wants something	…what it could want
… here's the CAT	… it looks like it sees something	…what it could see
… here's DADDY	… he's building, for some reason	…why Daddy is building
… here's DADDY	… he's hiding, for some reason	…why Daddy is hiding
… here's the CAT	… it's kissing something	…what it is kissing
… here's the CAT	… it's wearing something	…what it is wearing

License

Face time: Effects of shyness and attention to faces on early word learning

Matt Hilton
Maastricht University, the Netherlands
Lancaster University, UK

Katherine E. Twomey
University of Manchester, UK

Gert Westermann
Lancaster University, UK

Abstract: Previous research has shown that shyness affects children's attention during the fast-mapping of novel words via disambiguation. The current study examined whether shyness also affects children's attention when eye-gaze cues to novel word meanings are present. 20- to 26-month-old children's (N = 31) gaze was recorded as they viewed videos in which an onscreen actor sat at a table on which one novel and two familiar objects appeared. The actor looked at and labeled one of the objects, using a novel word if the target object was novel. Overall, shyness was associated with a stronger preference for looking at the actor's face, and less time looking at the object being labeled. These effects did not differ when the target object was novel or familiar, suggesting that shyness is related to attentional differences during object labeling generally, rather than specific processes involved in the disambiguation of novel words. No evidence was found of a relation between retention and shyness or attention during labeling.

Keywords: temperament; disambiguation; retention; early childhood; eye-tracking

Corresponding author: Matt Hilton, Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Oxfordlaan 55, Maastricht, 6229 EV, the Netherlands. Email: matt.hilton@maastrichtuniversity.nl

ORCID ID: https://orcid.org/0000-0002-6792-2208

Citation: Hilton, M., Twomey, K.T., & Westermann, G. (2023). Face time: Effects of shyness and attention to faces on early word learning. *Language Development Research, 3*(1), 156–181. https://doi.org/10.34842/2023.652

Introduction

During early language development, children are often required to determine word meanings before these meanings can be learned. This first stage of word learning is challenging because of referential uncertainty: There are multiple potential referents for any newly-encountered word, and there exists no reliably unambiguous cue as to the intended referent (Quine, 1960). Children must therefore become skilled at quickly disambiguating the meaning of unfamiliar words, and they make use of a range of different cues to do so. For example, a parent might look in the direction of a bowl containing a pink-colored bobbled fruit that their child has never seen before, and an array of familiar fruits. When the parent then says the novel word "lychee", their child will typically map this novel word to the unfamiliar fruit, rather than one of the familiar fruits (Halberda, 2006; Mervis & Bertrand, 1994), a behavior known as "fast-mapping" via disambiguation (Carey, 1978; Carey & Bartlett, 1978). Subsequent work has suggested that children can fast-map novel words via disambiguation as early as 17 months of age (Halberda, 2003; but see also Kucker et al., 2018), and by 24 months of age this behavior is reliably demonstrated in lab-based tasks (e.g., Axelsson et al., 2012; Bion et al., 2013; Horst & Samuelson, 2008; Horst et al., 2010). In this example, however, the child need not rely solely on fast-mapping cues to disambiguate the word "lychee": The child might also be close enough to notice that the parent is looking at the pink-colored fruit. From as early at 3 months of age, babies will shift their attention to look at the target object of somebody else's gaze (e.g., Hood et al., 1998), and from around 18 months of age, children can map a novel word to an object that is being looked and pointed at (Baldwin, 1993).

Experiments that examine children's novel word disambiguation typically find variability in children's performance. For example, in typical tests of fast-mapping via disambiguation, 24-month-old children incorrectly select a familiar object as the referent of a novel word on approximately one quarter of trials (e.g., Axelsson et al., 2012; Horst & Samuelson, 2008). Although children's incorrect selections on these tasks are often (implicitly) treated as reflective of random error (for example due to a passing distraction) or preference for a particular familiar object (non-linguistic preference effects; e.g., Moore et al., 1999), more recent evidence has indicated that these errors may reflect stable individual differences in children's novel word disambiguation. Hilton and Westermann (2017) examined whether fast-mapping errors can in part be explained by enduring temperament-based individual differences, rather than in-the-moment random error: They presented 24-month-old children with a typical experimental task and examined whether children's shyness could explain differences in their novel word disambiguation. Shyness is a biologically-based and enduring temperamental trait characterized as discomfort in (predominantly novel) social situations (Putnam et al., 2006), and it has been shown to affect children's vocabulary growth (Smith Watts et al., 2014; Spere et al., 2004). Hilton and Westermann found that when presented with an array of one novel and two familiar objects, shyer children were less likely to select

the novel object as the referent of the novel word (e.g., in response to the question *where's the koba?*) than less-shy children, indicating that children's fast-mapping via disambiguation is modulated by enduring temperament-related individual differences. Given the face-to-face nature of the task, however, it could not be concluded whether shyer children are generally less likely to fast-map via disambiguation, or whether their reduced fast-mapping was reflective of an unwillingness to engage with the task given the discomfort they felt in the novel social situation.

More recent work has attempted to probe the mechanism underlying the relation between shyness and disambiguation. For example, Hilton et al. (2019) removed some of the social demands of the word-learning task by converting it to a looking-while-listening task (e.g., Fernald et al., 1998): 20- to 26-month-old children sat on their parent's lap in a testing booth and viewed images of sets of three objects (one novel, two familiar) on a screen, while a familiar or the novel object was labeled via pre-recorded sentences played through speakers. Even though children were not required to offer a response on this task, shyness modulated their looking patterns across the array of objects during labeling. Specifically, shyness was linked to a reduction in attention to the target object, regardless of whether the heard label was novel or familiar. This finding indicates that shyer children's reduced fast-mapping could be explained in terms of differences in attentional processing. In a version of the same task with 30-month-old children, Axelsson et al. (2022) asked children to point to the referent of the heard word. It was found that children's approachability, a temperamental sub-domain tightly aligned with shyness, was negatively correlated with their pointing accuracy, meaning that shyer children were less likely to correctly select the novel object as the referent of a novel label, specifically on the second trial on which each novel word-object mapping was presented. Interestingly, results of this study also revealed a negative relation between children's temperamental reactivity, defined as children's tendency to feel intense emotions particularly in new contexts, and their looking time to both novel and familiar objects during labeling.

A question arising from this previous work is whether effects of shyness on fast-mapping are specific to disambiguation, or whether they come to bear on children's formation of novel word-object mappings more generally. Given shyer children's specific aversion to unfamiliar people (Putnam et al., 2006), it is plausible that shyness can modulate the formation of novel word-object mappings via social-based cues, such as eye-gaze provided by an unfamiliar adult. Somewhat counterintuitively, shyness is related to greater attention to faces, and in particular to eyes (Brunet et al., 2009; Matsuda et al., 2013; Wieser et al., 2009). These findings have typically been explained in terms of shyer children's hypervigilance to threatening or aversive stimuli. In general, children and adults show more rapid orientation and greater overall attention to threatening or aversive stimuli (e.g., Field, 2006), which is enhanced in shyer individuals who find novel social encounters aversive (e.g., Poole & Schmidt, 2021). In examining the role of shyness in children's formation of novel word-object mappings via eye-gaze cues, it

is important to establish whether shyness is related to hypervigilance to faces already during the second and third years of life, the time during which eye-gaze cues play an increasing role in children's word-object mapping (Baldwin, 1993).

It is, however, not immediately clear how a hypervigilance to faces would affect children's fast-mapping via eye-gaze. On the one hand, shyer children's increased attention to faces could support a more rapid and more accurate use of eye-gaze to determine the referent of a spoken word. On the other hand, children's general tendency to look longer at stimuli that they find aversive could mean that shyer children do not follow the eye-gaze cues to the referent, and therefore fail to map the word to the target object. In particular, it would be fruitful to examine these potential effects in the context of novel word disambiguation, given previous interpretations that shyer children's aversion to novel objects (e.g., Kagan et al., 1987; Rothbart, 1988) can explain their reduced target object selection and looking on fast-mapping tasks (Axelsson et al., 2022; Hilton & Westermann, 2017; Hilton et al., 2019). By examining the effect of shyness on looking times to the target object and the face during labeling, we will be better able to understand how aversion to novelty as a marker of shyness comes to bear on division of attention across novel faces and objects.

The formation of novel word-object mappings is, however, only the first stage of word learning. The newly-formed mappings must subsequently be retained, and recent work has argued that attentional processes during the disambiguation of a novel word are critical in determining whether the child will successfully retain the word-object mapping. For example, disambiguation alone is not sufficient to support retention of the newly-formed word-object mapping by 24-month-old children. Instead, retention of the word-object mapping is only demonstrated if the child's attention to the object is heightened following disambiguation (for example, by lifting it up and away from any competitors) while the word is repeated (Horst & Samuelson, 2008). A potential explanation for this finding is that disambiguation is driven by attention to familiar competitors: In order to eventually map the novel word to the novel object, the familiar competitors must first be ruled out as potential referents. Eye-tracking data supports this explanation by showing that looking behavior during disambiguation is characterized by equal looking towards familiar competitors as to the novel target object (Hilton et al., 2019; Twomey et al., 2018). Taken together, these findings suggest that attention to familiar competitors is critical for successful fast-mapping, while heightened attention to the target object is critical for retention of the newly formed word-object mapping. Successful word learning is therefore the result of a complex balance of attention across objects during disambiguation. Based on evidence that shyness is also related to a reduced retention of recently-formed novel word-object mappings (Axelsson et al., 2022; Hilton & Westermann, 2017), it is plausible that shyness-related differences in attention distribution during the formation of novel word-object mappings can explain this effect.

The aim of the current study was therefore to examine whether the effects of shyness on children's novel word disambiguation persist when eye-gaze cues to word meaning are also present. 20- to 26-month-old children were tested on an adaption of the looking-while-listening study used by Hilton et al. (2019): Participants were presented with images of one novel and two familiar objects on a screen while their eye movements were measured by an eye-tracker. Critically, the images were accompanied by an onscreen actor who looked at the target object while labeling it. In line with previous work (Axelsson et al., 2012; Hilton et al., 2019; Horst & Samuelson, 2008; Ma et al., 2022; Twomey et al., 2018), we compared children's looking behavior when disambiguating a novel word with their looking behavior when presented with a known word-object mapping, by including trials on which one of the familiar objects was labeled. We then examined whether children's shyness, as measured by the parent-report Early Childhood Behavior Questionnaire (ECBQ; Putnam et al., 2006), was related to their looking across the face and objects during labeling.

Based on previous findings that shyer children are hypervigilant to faces and eyes (e.g., Matsuda et al., 2013), we predicted that shyness would be positively related to looking to the face, and that this increased looking to the face would reduce shyer children's overall looking time to the target object during labeling. However, despite this reduction in looking time to the target object, we speculated that shyer children's hypervigilance to the face could mean that they are more responsive to the eye-gaze cues, and that these cues may serve to focus shyer children's attention to the target object relative to the competitor objects. Given that increased attention to the target object during labeling is related to a greater likelihood of retaining the word-object mapping (Hilton et al., 2019), we also examined whether any effect of shyness on attention to the target during labeling was also related to later retention. If shyer children are more responsive to eye-gaze cues, then any related focus on the target object relative to the competitors could serve to boost retention. Conversely, attention to competitors to rule them out as potential referents is also critical in supporting retention (Halberda, 2006), meaning that it is also possible that any heightened focus on the target object relative to the competitors may serve to weaken retention.

Method

Participants

A total of 31 typically developing children aged 20 and 26 months old took part in the study. All children were typically-developing monolinguals and were from predominantly white, middle-class families living in Lancaster, UK. There were 16 children in the 20-month age group (M = 20 m, 11 days; range = 19 m, 19 days to 20 m, 25 days; 7 girls) and 15 children in the 26-month age group (M = 26 m, 14 days; range = 25 m, 8 days to 27 m, 8 days; 4 girls). Data from an additional five 20-month-old children were excluded due to equipment error (n = 1), or because they were unable to adequately attend to the

experiment (e.g., due to distress or refusal to be in the testing suite; $n = 4$), and data from an additional four 26-month-old children were excluded due to equipment error ($n = 1$), because they were unable to adequately attend to the experiment (e.g., due to distress or refusal to be in the testing suite; $n = 2$), or because they showed no looking to the face on every disambiguation trial ($n = 1$). Families were recruited by contacting parents who had previously indicated interest in participating in child development research. Parents' travel expenses were reimbursed, and children were offered a gift of a storybook for participating.

Prior to their visit to the lab parents were requested to complete the Oxford CDI vocabulary checklist (Hamilton et al., 2000) for their children. Some parents could not complete the questionnaire prior to their visit and were therefore asked to take the questionnaire home and mail it back within a week of their visit. Questionnaire data for one 26-month-old child were missing due to the parent not returning the questionnaire and were replaced by the mean. The 20-month-old group had a mean productive vocabulary of 107 words (range = 7-413 words) and a mean receptive vocabulary of 245 words (range = 45-414 words). The 26-month-old group had a mean productive vocabulary of 246 words (range = 58-368 words) and a mean receptive vocabulary of 350 words (range = 232-414 words). As expected, the 26-month-old group had larger receptive and productive vocabularies than the 20-month-old group (receptive: t (29) = 3.32, p = .002; productive: t (29) = 3.62, p = .001; two-tailed).

Stimuli and design

Each child took part in disambiguation trials, which were presented on a computer screen, and retention trials, which involved the child selecting 3D objects from a tray. Visual stimuli for disambiguation trials consisted of digital photographs of eight objects selected because they are familiar to two-year-old children (ball, boat, car, cup, fork, motorbike, cell phone, shoe) and four novel objects (e.g., a plastic hand massager; see Figure 1). Each picture was of a similar size (approx. 70 x 70 mm) onscreen. Each novel object was assigned one of four novel pseudowords (cheem, koba, sprock, tannin), all of which were plausible English pseudowords and used in previous research (e.g. Hilton et al., 2019). Sixteen randomization orders were created, and each child in both age groups was assigned one of these sixteen orders. Within each order, the objects were randomly grouped into sets of three, with each set consisting of one novel object and two familiar objects. On subsequent retention trials children were presented with 3D objects, in line with previous work examining similar research questions (e.g. Zosh et al., 2013). Stimuli for the warm-up trials consisted of three familiar objects (helicopter, rubber duck, fork), and the novel objects that had been seen during disambiguation were used for the four retention trials. These objects were all of a similar size (approx. 95 x 70 x 50 mm).

A separate video was created for each of the 12 disambiguation trials. Each video began

Figure 1. ***Photographs of objects used in the study. Panel a) shows example familiar stimuli shown on disambiguation trials. Panel b) shows the four novel stimuli.***

by showing an unfamiliar female Caucasian actor in her mid-20s sitting at a table and looking with a neutral expression at the camera. After 600 ms, pictures of three objects then bounced simultaneously onto the table from the bottom of the screen, one to the left, one to the middle, and one to the right (see Figure 2 for an example still) . The actor then looked at the target object (approx. 3000 ms after the trial onset) and labeled it three times in a neutral tone, embedded within a consistent script (Look, it's a _____! Can you see the _____? Wow, it's a _____!). Each sentence was produced and recorded as in real-time, meaning that the precise onset of individual labels varied slightly across trials. After she had finished speaking (approx. 10,400 ms after she began), the actor looked back at the camera with a neutral expression, and the objects disappeared. Each set was presented three times. On the first two presentations, the novel object acted as the target (novel label trials), and on the final presentation a randomly selected familiar object acted as the target (familiar label trials), to ensure that on novel label trials the novel object had not previously been seen on a preceding familiar trial. The order in which the sets were presented was pseudorandomized, with the constraint that no set was presented more than twice successively. Across the three presentations of a given set, the target appeared once on the left, once in the middle and once on the right.

Figure 2. ***Example video still from disambiguation trial.***

Procedure

Shyness questionnaire

Parents completed the shyness scale of the Early Childhood Behavior Questionnaire (ECBQ; Putnam et al., 2006) during their visit. In order to reduce demand biases in parents' responses, three other unrelated questions taken from the ECBQ were included within the questionnaire, but these responses were not analyzed. Presenting questions relating to only these two subscales avoided overburdening parents with questionnaires, and such a procedure is in line with previous work using temperament questionnaires (Justice et al., 2008; Rudasill et al., 2014; Spere & Evans, 2009). The ECBQ is a standardized parent report measure of 18- to 36-month-old children's emerging temperament. Twelve items measure the child's shyness, and each item asks parents to rate from 1-7 (1 = never, 7 = always) how often over the past two weeks their child has demonstrated shy-type behaviors (e.g., "when playing with unfamiliar children, how often did your child seem uncomfortable?"). Averaging across the 12 questions (Cronbach's $\alpha = .84$)

yields a score for each child between 1 (not at all shy) and 7 (extremely shy).

Disambiguation trials

Children sat on their parent's lap approximately 60 cm from a computer monitor mounted above a Tobii x120 eye-tracker, which recorded children's gaze data from both eyes at a sampling rate of 60 Hz. Parents were instructed not to speak to their child or look at the screen during stimulus presentation to avoid influencing their child's looking behavior. The experimenter monitored the session via a video camera to ensure that these instructions were complied with. Videos were imported into Tobii Studio (version 3.4) and programmed to run sequentially. Before stimuli were presented, the gaze of each child was calibrated using a five-point procedure: A colorful child-friendly animation (e.g., a wobbling duck) was displayed in the four corners and middle of a 3x3 grid, and calibration accuracy was checked and repeated if necessary. Disambiguation trials followed immediately after calibration.

After every fourth trial, a four-second long child-friendly animation accompanied by an exciting sound effect (e.g., rattling sounds) was displayed in order to keep children's attention on the screen. After disambiguation trials were completed, children took a five-minute break during which they played in an adjacent room. This break was included in line with previous work (e.g., Horst & Samuelson, 2008) to ensure that the subsequent retention phase required recall from long-term memory.

Data coding and cleaning. The raw data files were exported from Tobii Studio (version 3.4) and processed in R (version 4.2.1; R Core Team, 2022) via R Studio (version 1.2.5001; RStudio Team, 2020) with the tidyverse package (version 1.3.2; Wickham et al., 2019). For each participant, the data file showed a timestamp for each data sample and the corresponding x-y coordinates of the child's gaze on the screen. Four square object Areas of Interest (AOIs) were created in the areas of the screen where the stimuli were displayed. All AOIs measured 400 by 400 pixels. An object AOI covered each position in which the objects appeared: left, middle and right. There was a gap of 100 pixels between object AOIs. A margin of 20 pixels separated AOIs from the left and right edge of the screen, and a margin of 40 pixels separated the AOIs from the bottom of the screen. A further AOI covering the position of the actor's face measuring 400 by 400 pixels was defined. AOIs did not overlap. Continuous gaze within an AOI was counted as a fixation. If continuous gaze within an AOI was interrupted for less than 60 ms, this interruption was recoded as a continuation of that fixation, because this was most likely due to blinking or eye-tracking errors rather than the child rapidly re-orienting their attention (0.16 % of data samples were recoded in this way). Only data samples collected following the approximate onset of the target name and until the disappearance of the objects (4500 – 10400 ms from video onset) were analyzed. The proportion of looking time in each AOI was calculated for each trial by dividing

the sum of gaze samples in the AOI by the sum of gaze samples that fell into any AOI[1], and these proportions were converted to two log-gaze proportion ratios for analysis (Arai et al., 2007; Borovsky et al., 2016). Proportions of 0 were transformed to 0.01 to allow for log transformation. A face-vs-target log-gaze proportion ratio was calculated by log transforming the proportion looking time to the face divided by proportion looking time to the target object, log(*P*[Face]/*P*[Target]). A log-gaze proportion ratio of zero reflects equal looking across face and target, a positive log-gaze proportion ratio reflects preferential looking to the face, and a negative log-gaze proportion ratio reflects preferential looking to the target object. The magnitude of the log-gaze proportion ratio reflects the strength of the preference. A target-vs-competitors log-gaze proportion ratio was also calculated, log(*P*[Target]/*P*[Competitors]).

Retention Trials

Following the five-minute break, children returned to the testing room to take part in retention trials. Retention trials began with a warm-up task. Children were seated on their parent's lap opposite the experimenter. The experimenter presented the three familiar objects to the child on a tray specially divided into three sections, initially out of reach of the child, for approximately three seconds. Children were then asked to select one of the objects (e.g., *Where's the duck?*), the tray was pushed forward into the child's reach, and their response was recorded. If the child selected the correct object, both the experimenter and parent praised the child, or if the child selected an incorrect object, the experimenter and parent encouraged the child to select the correct one. If the child failed to respond after two further prompts, the tray was removed, and the next trial began. On each subsequent trial, the three objects were rearranged out of sight of the child, and children were asked for a different object. The warm-up task continued until children had selected the correct object three times in a row. Across the warm-up task, each target object appeared in each section of the tray at least once. Retention trials continued in the same manner as the warm-up task, with two differences. First, no praise or encouragement was offered following the child's response: the experimenter simply replied with a neutral *thank you*. Second, retention trials consisted of the novel objects seen during the disambiguation phase. On each trial the child was presented with a target alongside two other randomly selected novel objects from the disambiguation trials and was asked for the target using the appropriate novel word. There was one retention trial for each novel object; each child therefore participated in four retention trials. The order of retention trials was randomly determined, as was the location of the target on the tray.

[1] e.g., $P[\text{Face}] = \frac{\text{sum samples [Face]}}{\text{sum samples [Face]}+\text{sum samples [Target]}+\text{sum samples [Competitors]}}$

Results

Proportional looking during disambiguation

To determine whether shyness was associated with increased looking to the face relative to the target object, the face-vs-target log-gaze proportion ratios were submitted to a linear mixed effects model (LMEM) with main effects of shyness score (mean-centered across all models) and trial type (sum coded: familiar label trial = 1, novel label trial = -1 across all models), with their interaction, by-participant correlated random slopes for shyness score and intercepts and by-target random intercepts[2]. Log-gaze proportion ratios from six trials were excluded from the analysis because the child looked at neither the face nor the target object. Results are shown in Table 1.

Table 1. ***Results of linear mixed effects model for face-vs-target log-gaze proportion ratios. Significant predictors are highlighted in bold.***

	β	*SE*	*t*	χ^2	*df*	*p*
intercept	0.83	0.22	3.76			
Trial Type	-0.35	0.13	-2.68	5.38	1	.020
Shyness	1.04	0.32	3.28	6.74	1	.009
Trial Type x Shyness	0.14	0.15	0.94	0.87	1	.350

For this analysis, a positive log-gaze proportion ratio indicates preferential looking to the face relative to the target object, and the positive intercept estimate therefore indicates that overall, children looked more to the face than to the target object. Interestingly, the significant negative main effect of trial type indicates that children showed a greater preference to look at the target object relative to the face on familiar label trials (*M* log-gaze proportion ratio = 0.45, *SD* = 1.54) than on novel label trials (*M* = 1.13, *SD* = 1.44). Critically, the significant main effect of shyness indicates that shyness was associated with greater tendency to look at the face, and less looking to the target object. These findings confirm the prediction that shyer children would look more to the face relative to the target object.

Axelsson et al. (2022) found that the relation between approachability and looking behavior differed between the first and second labeling event of the novel object. In the current study, children also saw each novel object labeled on two separate trials.

[2]All linear mixed effects models were conducted using the lme4 package (version 1.1-30 Bates et al., 2015) and initially defined with maximal random effects structures, which were then simplified until convergence (D. J. Barr et al., 2013). *p*-values for fixed effects were obtained using sequential likelihood ratio tests, and *p*-values from follow-up pairwise comparisons were Bonferroni-corrected unless otherwise stated. Initial analyses revealed no effect of age on looking during disambiguation, so age was excluded as a fixed factor in models of disambiguation to maximize power (cf. Hilton et al., 2019). Estimated random effect variances and R formulae for each model are reported in the supplementary materials.

We therefore ran a subsequent exploratory analysis examining face-vs-target log-gaze proportion ratio on novel label trials only, including fixed factors of shyness and labeling event (sum coded: first labeling event =-1, second labeling event = 1). As expected, analyses confirmed a significant fixed effect of shyness (β = 0.99, *SE* = 0.33, *t* = 3.00, χ^2 = 6.69, *p* = .010). The fixed effect of labeling event was marginally non-significant (β = -0.22, *SE* = 0.12, *t* = -1.81, χ^2 = 3.66, *p* = .056), suggesting that children tended to look less to the face relative to the target object on the second novel labeling event than the first. Critically, however, no interaction between shyness and labeling event was found (β = 0.19, *SE* = 0.17, *t* = 1.14, χ^2 = 1.29, *p* = .260).

The finding that shyer children attended more to the face relative to the target object does not rule out the possibility that these children were more responsive to the eye-gaze cues to disambiguate the heard label. More accurate use of the eye-gaze cues could be reflected by greater attention to the target object relative to the competitor objects. Target-vs-competitor log-gaze proportion ratios were therefore submitted to a LMEM with main effects of shyness score and trial type with their interaction, by-participant correlated random slopes for trial type and intercepts and by-target random intercepts. Log-gaze proportion ratios from 24 trials were excluded from the analysis because the child looked at neither the target object nor the competitor objects. Results can be seen in Table 2.

Table 2. ***Results of linear mixed effects model for target-vs-competitor log-gaze proportion ratios. Significant predictors are highlighted in bold.***

	β	*SE*	*t*	χ^2	*df*	*p*
intercept	0.01	0.17	0.05			
Trial Type	0.34	0.14	2.45	5.49	1	.019
Shyness	-0.44	0.25	-1.80	2.53	1	.110
Trial Type x Shyness	-0.16	0.20	-0.80	0.64	1	.420

For this analysis, a positive log-gaze proportion ratio reflects preferential looking to the target relative to the competitors. The intercept estimate therefore indicates that children overall looked roughly equally across the target object and the competitor objects. The main effect of trial type indicates that children looked preferentially to the target object on familiar label trials (*M* log-gaze proportion ratio = 0.31, *SD* = 1.55) but to the competitors on novel label trials (*M* = -0.30, *SD* = 0.98). These findings suggest that, despite the presence of eye-gaze cues to the target object on all trials, attention to the target was heightened relative to competitors only on familiar label trials. On novel label trials, children still sought to rule out familiar competitors as potential referents. It therefore appears that the eye-gaze cues did not override children's fast-mapping via disambiguation behaviors. We will return to this point in the discussion. Furthermore, no main effect of shyness was found, providing no evidence that shyer children's increased attention to the face served to focus their attention more on the target object relative to the competitors.

We also examined whether the relation between shyness and target-vs-competitor log-gaze proportion ratio differed between the first and second labeling event of the novel object. We therefore ran a further LMEM on data from novel label trials, with main effects of shyness and labeling event (sum coded: first labeling event =-1, second labeling event = 1). These analyses revealed no effect of shyness (β = -0.23, *SE* = 0.32, *t* = -0.71, χ^2 = 0.61, *p* = .43) and no effect of labeling event (β = -0.10, *SE* = 0.12, *t* = -0.86, χ^2 = 0.33, *p* = .57). There was also no interaction between shyness and labeling event (β = -0.27, *SE* = 0.18, *t* = -1.53, χ^2 = 2.30, *p* = .13).

Looking time during disambiguation

The analysis of log-gaze proportion data revealed that shyness modulated children's division of attention across the face and target object during labeling. While this analysis revealed that increased shyness was associated with greater attention to the face relative to the target object, it was unclear whether this effect was related to reduced looking times to the target object. For example, if a highly-attentive child shows a stronger preference for the face relative to the target object, they could still spend longer looking at the target object than a child with a weaker preference for the face and reduced overall looking. A further series of LMEMs were therefore run in order to examine whether differences in children's division of attention, as measured by log-gaze proportion ratios, affected summed looking times to the different AOIs onscreen. Looking time on each trial in seconds was submitted to a LMEM with main effects of shyness score, trial type, and AOI hit type (sum coded: contrast 1: competitor = 1, face = 0, target = -1; contrast 2: competitor = 0, face = 1, target = -1) with their interactions and by-participant random intercepts. Results can be seen in Table 3.

Table 3. ***Results of linear mixed effects model for looking time during disambiguation. Significant predictors are highlighted in bold.***

		β	*SE*	*t*	χ^2	*df*	*p*
intercept		1.56	0.06	27.16			
AOI Hit Type	contrast 1	-0.47	0.06	-8.61	211.01	2	< .001
	contrast 2	-0.79	0.06	14.41			
Trial Type		-0.11	0.04	-2.79	7.16	1	.007
Shyness		0.02	0.08	0.02	<0.01	1	.950
AOI Hit Type x Shyness	contrast 1	-0.17	0.08	-2.24	61.88	2	< .001
	contrast 2	-0.56	0.08	7.22			
AOI Hit Type x Trial Type	contrast 1	-0.07	0.06	-1.21	27.33	2	< .001
	contrast 2	-0.21	0.06	-3.81			
Shyness x Trial Type		0.02	0.06	0.40	0.16	1	.691
3-way Interaction	contrast 1	0.09	0.08	1.14	1.29	2	.524
	contrast 2	-0.05	0.08	-0.62	1.29		

The LMEMs revealed a main effect of AOI hit type. As expected based on results of the

proportional looking time data, this effect was due to longer looking times to the face (M = 2.41 s; SD = 0.99) than to the target object (M = 1.21 s; SD = 0.68; $\chi^2(1) = 135.29$, $p < .001$) and longer looking time to the face than to the competitor objects (M = 1.16 s; SD = 0.47; $\chi^2(1) = 144.29$, $p < .001$), while looking time to competitors and target did not differ ($\chi^2(1) = 0.48$, $p > .99$).

The main effect of trial type reveals that children looked generally longer to all AOIs on novel label trials (M = 1.67 s; SD = 0.27) than on familiar label trials (M = 1.46; SD = 0.53). There are several potential explanations for this effect. For example, novel label trials were likely more cognitively demanding and therefore required greater attention to determine the correct referent. The experimental design may also explain this effect: The familiar label trials were designed to appear after the child had already seen the novel object labeled twice, so that on familiar label trials participants may have been less attentive due to experimental fatigue.

Critically, an interaction between shyness and AOI hit type was revealed, showing that shyness modulated children's looking times to the three AOI hit types. However, follow-up analysis of simple main effects did not reveal a significant relation between shyness and looking time to the face ($\chi^2(1) = 5.02$, $p = .08$), target ($\chi^2(1) = 5.14$, $p = .07$) or competitors ($\chi^2(1) = 2.87$, $p = .27$). Although not originally planned, we also opted to re-examine these effects using an alternative p-value adjustment to better understand which effects might be driving the significant interaction. The Benjamini-Hochberg (Benjamini & Hochberg, 1995) adjustment, which in contrast to the Bonferroni correction attempts to control for the false-discovery rate, indicated that shyness was positively related to looking times to the face ($p = 0.038$), negatively related to looking time to the target ($p = 0.038$), and unrelated to looking time to the competitors ($p = 0.090$). Taken together with the proportional analyses, we therefore tentatively conclude that shyness was associated with a decrease in looking time to the target and an increase in looking time to the face (see Figure 3).

Finally, pairwise comparisons of the interaction between AOI hit type and trial type (see Figure 4) revealed that children looked longer to the face on novel label trials (M = 2.61 s, SD = 1.04) than on familiar label trials (M = 2.00 s, SD = 1.14; $\chi^2(1) = 17.06$, $p < .001$), and they also looked longer to the competitors on novel label trials (M = 1.29 s, SD = 0.58) than on familiar label trials (M = 0.94 s, SD = 0.56; $\chi^2(1) = 11.70$, $p = .002$). These findings indicate that, when the heard label was novel, children's disambiguation was marked by greater looking to the face and the competitor objects. Conversely, children looked longer to the target on familiar label trials (M = 1.42 s; SD = 0.97) than on novel label trials (M = 1.11 s, SD = 0.62; $\chi^2(1) = 9.55$, $p = .006$).

We also examined whether the relation between shyness and looking times to the AOIs differed across first and second labeling events of the novel objects. We therefore ran LMEMs on data from the novel label trials only. Due to convergence-related issues,

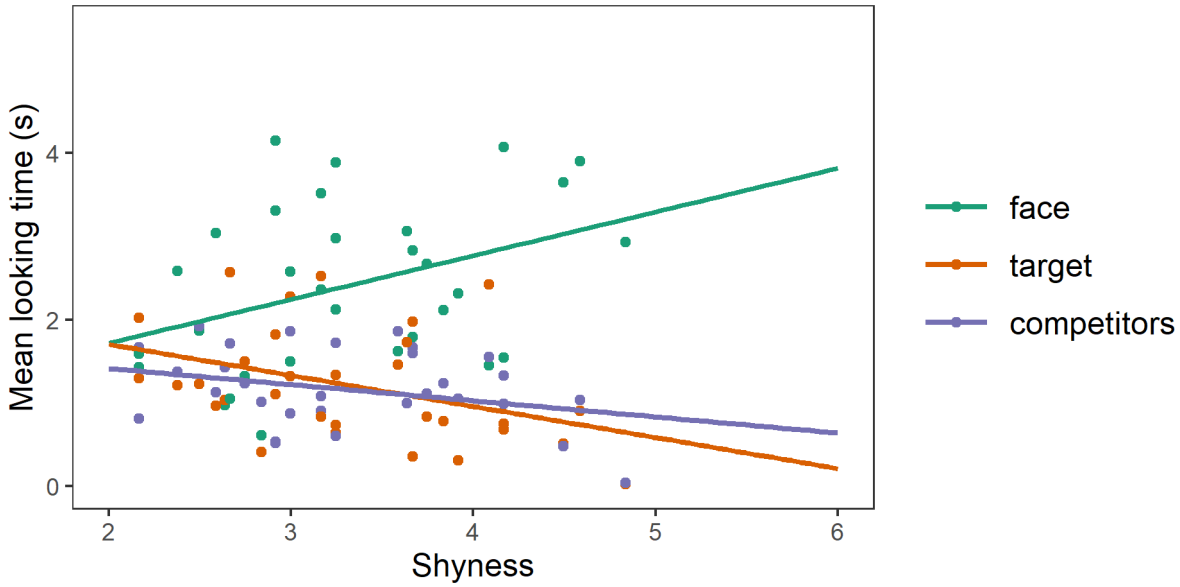


Figure 3. ***Children's shyness scores plotted against their mean looking to each AOI during disambiguation. For illustration, lines are linear regressions.***

three separate LMEMs were run, with summed looking time in each AOI (face, target, competitors) as dependent variable, and shyness, labeling event (sum coded: first labeling event =-1, second labeling event = 1), and their interaction as fixed factors. Full results can be found in the supplementary materials. No significant interaction between shyness and labeling event were found, suggesting that in our study, the relation between shyness and looking time was not modulated by whether the child has already seen the novel object labeled on a previous trial.

Retention Trials

Four children in the 20-month group and one child in the 26-month group did not complete training, and so were excluded from retention analyses. Retention trials were scored 1 if the child selected the correct referent and 0 if they did not. In order to test whether children demonstrated retention above levels expected by chance, a proportion correct retention score was calculated for each child and submitted to a one-sample *t*-test with chance set at 0.33. The 20-month-old group did not demonstrate retention above levels expected by chance alone (M = .38, SD = .27, t (11) = 0.57, p = .578). The 26-month-old group also showed no evidence of retaining the novel label meanings (M = .39, SD = .19, t (13) = 1.24, p = .235). While these analyses reveal that overall children did not retain the label-object associations that were presented during disambiguation, in line with Hilton et al. (2019), we next examined whether shyness and proportional target looking during disambiguation predicted retention scores. Trial-by-trial retention scores (correct = 1, incorrect = 0) were submitted to a binomial generalized LMEM with

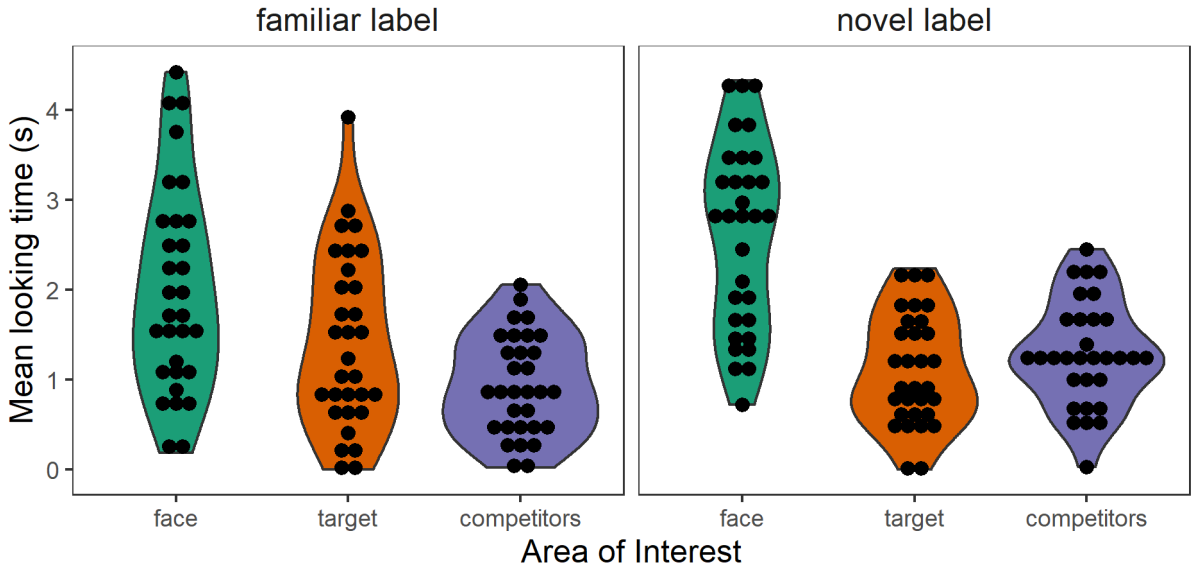


Figure 4. ***Mean looking times in seconds for each participant to each area of interest during disambiguation. The left panel depicts looking on familiar label trials, and the right panel depicts looking on novel label trials.***

main effects of shyness score and target-vs-competitor log-gaze proportion ratio, their interaction, and correlated by-participant random slopes and intercepts for log-gaze proportion ratios and uncorrelated by-target random intercepts and slopes for log-gaze proportion ratios. Results are presented in Table 4, and reveal no significant main effects or interactions, providing no evidence that children's shyness and looking during disambiguation were related to their retention of the label-object associations

Table 4. ***Results of linear mixed effects model for retention trials.***

	β	*SE*	*t*	χ^2	*df*	*p*
intercept	-0.79	0.45	-1.74			
Target/competitor looking[3]	-0.02	0.18	-0.12	0.15	1	.700
Shyness	-0.32	0.33	-0.95	1.48	1	.224
Target/competitor looking x shyness	0.16	0.20	0.81	0.59	1	.444

In order to examine whether looking times to the target or face were significantly related to retention, a further series of LMEMs were run, using summed looking time to the target or face during disambiguation as a main effect alongside shyness score and disambiguation trial type (first novel labeling event vs. second novel labeling event). Full details of model specification and results can be found in the supplementary materials. These analyses also revealed no significant effects of, or interactions with, looking time

[3]Target vs. competitor log-gaze proportion

during disambiguation. These findings therefore provide no evidence that looking behavior during disambiguation is related to retention of the word-object mappings. It is possible that children's poor retention was due to their inability to transfer learning from the screen-based disambiguation task to the live-3D retention task. On the other hand, it is also possible that the presence of the onscreen actor reduced children's overall attention to the target so that they could not sufficiently encode the label-object association. We will return to this point in the discussion.

Discussion

The current work examined whether shyness modulated children's attentional processing during novel word disambiguation, when both eye-gaze and disambiguation cues are provided. The findings suggest that at 20 to 26 months of age, shyness as measured by the ECBQ is related to heightened attention to faces. Critically, this heightened attention to the face did not confer an advantage on shyer children in interpreting eye-gaze cues: shyer children showed the same pattern of looking across target and competitor objects as less-shy children. Instead, the findings indicated that shyer children's heightened attention to the face during labeling reduced their looking time to the target object, which could have weakened their encoding of the word-object mapping. These results could also explain previous findings showing shyer children's reduced novel word disambiguation and retention when measured on a typical face-to-face lab task (Hilton & Westermann, 2017). The current study, however, found no evidence that looking behavior differed with a repeated exposure to the novel word-object mapping, nor that shyness or looking during novel word disambiguation were related to retention of these novel word-object mappings.

The findings that shyer children showed a stronger preference to look at the face, that this preference likely resulted in a decrease in attention to the target object, and that looking to the target object did not differ relative to competitor object looking, indicate that shyer children may struggle in word learning tasks because they do not encode the target object sufficiently to form a robust word-object mapping during disambiguation. Previous work has concluded that the formation of a label-object association during disambiguation is the product of a complex balance of attention to all available cues. For example, increased attention to a target object during disambiguation has been found to increase the likelihood that this word-object mapping will be retained (Axelsson et al., 2012; Hilton et al., 2019), although removing competitors from the task, despite increasing attention to the target during labeling, reduces retention (Zosh et al., 2013). Similarly, while children make use of eye-gaze cues to form label-object associations from as early as 15 months-of-age (Houston-Price et al., 2006), these cues do not improve learning of label-object associations by 18-month-old children if competitors are highly salient (Moore et al., 1999). The finding that shyness was associated with reduced looking to the target object during labeling replicates those of previous studies that presented images of the target and competitor objects on a blank background (i.e.,

no onscreen actor or social cues; Axelsson et al., 2022; Hilton et al., 2019). Critically, looking time to the target was reduced not just when the label meaning had to be disambiguated on novel label trials, but also on familiar label trials, when the label meaning was already known. It therefore appears that shyness is related to a general modulation of attentional processes during labeling, instead of individual differences specifically in disambiguation-related cognitive processes.

Our results are in line with previous findings that shyer individuals show heightened attention to eyes and social cues (Brunet et al., 2009; Matsuda et al., 2013). Critically, however, this effect was previously found in older children and in studies that presented stimuli containing only faces. The current study thus extends these previous findings by demonstrating that shyer children also attended preferentially to faces when alternative non-social stimuli (i.e., the target and competitor objects) were displayed. Previous work focusing on adults (Wieser et al., 2009) or older children (Brunet et al., 2009) has suggested that shyer individuals are hyper-vigilant to faces because of heightened self-consciousness (Crozier & Perkins, 2002), meaning that they are more attentive to any social signals that can be conveyed by other people's eyes, and this explanation might apply to the current study. An alternative, lower-level explanation for shyer children's preferential attention to faces in our study could be that it is driven by a formed association between unfamiliar people and feelings of anxiety, because we know that young children show an attentional bias to anxiety-inducing stimuli (e.g., pictures of snakes or angry faces; LoBue & DeLoache, 2010). Critically, shyer children's preferential attention to the face did not result in a more accurate use of eye-gaze cues to disambiguate the novel word: their division of attention across target and competitors on novel label trials did not differ from less-shy children's. This result suggests that increased face looking in shyer children serves to disrupt the attentional processes underlying novel word disambiguation.

Our findings raise some important issues for an understanding of word learning in general. First, despite the presence of eye-gaze cues, all children on average showed the same looking pattern as when objects were displayed on a blank background (Hilton et al., 2019): more looking to the competitor objects than to the target object on novel label trials, and greater attention to the target object than competitor objects only on familiar label trials. Previous work has indicated that even by 18 months of age, eye-gaze cues are not reliably attended to if competitors are highly salient (Moore et al., 1999). However, this same study found that by 24 months of age children will attend more to gaze-cued objects even in the presence of highly salient competitors, which is in contrast to the current study finding no difference in target object looking between 20- and 26-month-old children. Instead, it appears that in our study children did not capitalize on the eye-gaze cues to determine the referent of the novel label, but also disambiguated to eliminate competitors as potential referents. This finding is further evidence of the complex interplay between social and non-social cues to novel word disambiguation (e.g., Ma et al., 2022; MacDonald et al., 2017; Yurovsky & Frank, 2017).

Second, we found no relation between looking during disambiguation and children's retention of the new label-object associations. One possibility is that children did not transfer their learning from the 2D pictures of the objects to the actual 3D objects, known as the video deficit effect (Krcmar et al., 2007; Robb et al., 2009; see R. Barr, 2010, for a review), although children have shown no difficulty with this transfer in other studies (e.g., Zosh et al., 2013). Furthermore, despite previous evidence that heightened attention to the target object during disambiguation predicts successful retention of the word-object mapping (e.g., Axelsson et al., 2012), we found no association between looking to the target object and retention. Alongside the video deficit effect, a possible explanation for this finding could be that the presence of the onscreen actor reduced overall attention to the target object during disambiguation below levels sufficient to support retention.

Given that shyness in early childhood is marked specifically by inhibited behavior around unfamiliar adults and in unfamiliar settings (Putnam et al., 2006), it is possible that the shyness-related effects found in the current study are specific to the unfamiliar lab setting combined with the unfamiliarity of the onscreen actor. Follow-up studies could examine this aspect by manipulating the familiarity of the context or the person labeling the objects. The increased availability of mobile and head-mounted eye-tracking would allow, for example, for testing in the child's home or examining children's looking patterns when a familiar adult is labeling the object. These studies would also help us better understand how children's looking behavior on screen-based tasks, such as the current one, relate to children's learning in real-life settings.

Overall, this work shows that shyness exerts a robust effect on attention processing during novel word disambiguation. Specifically, our work demonstrates that the dynamic balance of attention to target object, competitor objects and eye-gaze cues during novel word disambiguation is modulated by shyness. While effects of shyness on social and emotional adjustment have been well-established (e.g., Coplan & Arbeau, 2008), the current study contributes to a growing body of literature that indicates that shyness modulates developing cognitive systems as well. Although shyness in early childhood does not appear to have long-term detrimental direct effects on later language abilities (e.g., Spere & Evans, 2009) or academic achievement (Hughes & Coplan, 2010; Zhang et al., 2017), this growing body of evidence suggests shyness is related to stable individual differences in cognitive processes involved in language development. By better understanding these individual differences, we can begin to support educators and practitioners in determining when children's differential behavior and development are due to normal shyness-related individual differences, or are indicative of more atypical development. Most importantly, work should now begin to further disentangle the dynamic relations between attentional processing, language development and shyness.

References

Arai, M., van Gompel, R. P., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology, 54*(3), 218–250. https://doi.org/https://doi.org/10.1016/j.cogpsych.2006.07.001

Axelsson, E., Churchley, K., & Horst, J. (2012). The right thing at the right time: Why ostensive naming facilitates word learning. *Frontiers in Psychology, 3*. https://doi.org/10.3389/fpsyg.2012.00088

Axelsson, E., Othman, N. N., & Kansal, N. (2022). Temperament and children's accuracy and attention during word learning. *Infant Behavior and Development, 69*, 101771. https://doi.org/https://doi.org/10.1016/j.infbeh.2022.101771

Baldwin, D. A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology, 29*, 832–843. https://doi.org/10.1037/0012-1649.29.5.832

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*(3), 255–278. https://doi.org/https://doi.org/10.1016/j.jml.2012.11.001

Barr, R. (2010). Transfer of learning between 2d and 3d sources during infancy: Informing theory and practice [Television and Toddlers: The Message, the Medium, and Their Impact on Early Cognitive Development]. *Developmental Review, 30*(2), 128–154. https://doi.org/https://doi.org/10.1016/j.dr.2010.03.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bion, R. A., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30months. *Cognition, 126*(1), 39–53. https://doi.org/https://doi.org/10.1016/j.cognition.2012.08.008

Borovsky, A., Ellis, E. M., Evans, J. L., & Elman, J. L. (2016). Semantic structure in vocabulary knowledge interacts with lexical and sentence processing in infancy. *Child Development*, *87*(6), 1893–1908. https://doi.org/https://doi.org/10.1111/cdev.12554

Brunet, P. M., Heisz, J. J., Mondloch, C. J., Shore, D. I., & Schmidt, L. A. (2009). Shyness and face scanning in children. *Journal of Anxiety Disorders*, *23*(7), 909–914. https://doi.org/https://doi.org/10.1016/j.janxdis.2009.05.009

Carey, S. (1978). Linguistic theory and psychological reality. In M. Halle, J. Bresnan, & A. Miller (Eds.). MIT Press.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford child language conference*, *15*, 17–29.

Coplan, R. J., & Arbeau, K. A. (2008). The stresses of a "brave new world": Shyness and school adjustment in kindergarten. *Journal of Research in Childhood Education*, *22*(4), 377–389. https://doi.org/10.1080/02568540809594634

Crozier, W. R., & Perkins, P. (2002). Shyness as a factor when assessing children. *Educational Psychology in Practice*, *18*(3), 239–244. https://doi.org/10.1080/0266736022000010267

Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*(3), 228–231. https://doi.org/10.1111/1467-9280.00044

Field, A. P. (2006). Is conditioning a useful framework for understanding the development and treatment of phobias? [Anxiety of childhood and adolescence: Challenges and opportunities]. *Clinical Psychology Review*, *26*(7), 857–875. https://doi.org/https://doi.org/10.1016/j.cpr.2005.05.010

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23–B34. https://doi.org/https://doi.org/10.1016/S0010-0277(02)00186-5

Halberda, J. (2006). Is this a dax which i see before me? use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cognitive Psychology*, *53*(4), 310–344. https://doi.org/https://doi.org/10.1016/j.cogpsych.2006.04.003

Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a british communicative development inventory. *Journal of Child Language*, *27*(3), 689–705. https://doi.org/10.1017/S0305000900004414

Hilton, M., Twomey, K. E., & Westermann, G. (2019). Taking their eye off the ball: How shyness affects children's attention during word learning. *Journal of Experimental Child Psychology*, *183*, 134–145. https://doi.org/https://doi.org/10.1016/j.jecp.2019.01.023

Hilton, M., & Westermann, G. (2017). The effect of shyness on children's formation and retention of novel word–object mappings. *Journal of Child Language*, *44*(6), 1394–1412. https://doi.org/10.1017/S030500091600057X

Hood, B. M., Willen, J. D., & Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science*, *9*(2), 131–134. https://doi.org/10.1111/1467-9280.00024

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128–157. https://doi.org/https://doi.org/10.1080/15250000701795598

Horst, J. S., Scott, E. J., & Pollard, J. A. (2010). The role of competition in word learning via referent selection. *Developmental Science*, *13*(5), 706–713. https://doi.org/https://doi.org/10.1111/j.1467-7687.2009.00926.x

Houston-Price, C., Plunkett, K., & Duffy, H. (2006). The use of social and salience cues in early word learning. *Journal of Experimental Child Psychology*, *95*(1), 27–55. https://doi.org/https://doi.org/10.1016/j.jecp.2006.03.006

Hughes, K., & Coplan, R. J. (2010). Exploring processes linking shyness and academic achievement in childhood. *School Psychology Quarterly*, *25*(4), 213–222. https://mu.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=2010-26288-003&site=ehost-live&scope=site

Justice, L. M., Cottone, E. A., Mashburn, A., & Rimm-Kaufman, S. E. (2008). Relationships between teachers and preschoolers who are at risk: Contribution of children's language skills, temperamentally based attributes, and gender. *Early Education and Development*, *19*(4), 600–621. https://doi.org/10.1080/10409280802231021

Kagan, J., Reznick, J. S., & Snidman, N. (1987). The physiology and psychology of behavioral inhibition in children. *Child Development*, *58*(6), 1459–1473. Retrieved February 8, 2023, from http://www.jstor.org/stable/1130685

Krcmar, M., Grela, B., & Lin, K. (2007). Can toddlers learn vocabulary from television? an experimental approach. *Media Psychology*, *10*(1), 41–63. https://doi.org/10.1080/15213260701300931

Kucker, S. C., McMurray, B., & Samuelson, L. K. (2018). Too much of a good thing: How novelty biases and vocabulary influence known and novel referent selection in 18-month-old children and associative learning models. *Cognitive Science, 42*(S2), 463–493. https://doi.org/https://doi.org/10.1111/cogs.12610

LoBue, V., & DeLoache, J. S. (2010). Superior detection of threat-relevant stimuli in infancy. *Developmental Science*, *13*(1), 221–228. https://doi.org/https://doi.org/10.1111/j.1467-7687.2009.00872.x

Ma, L., Twomey, K., & Westermann, G. (2022). The impact of perceived emotions on toddlers' word learning. *Child Development*, *93*(5), 1584–1600. https://doi.org/https://doi.org/10.1111/cdev.13799

MacDonald, K., Yurovsky, D., & Frank, M. C. (2017). Social cues modulate the representations underlying cross-situational learning. *Cognitive Psychology, 94*, 67–84. https://doi.org/https://doi.org/10.1016/j.cogpsych.2017.02.003

Matsuda, Y.-T., Okanoya, K., & Myowa-Yamakoshi, M. (2013). Shyness in early infancy: Approach-avoidance conflicts in temperament and hypersensitivity to eyes during initial gazes to faces. *PLOS ONE*, *8*(6), 1–7. https://doi.org/10.1371/journal.pone.0065476

Mervis, C. B., & Bertrand, J. (1994). Acquisition of the novel name-nameless category (n3c) principle. *Child Development, 65*(6), 1646–1662. Retrieved October 24, 2022, from http://www.jstor.org/stable/1131285

Moore, C., Angelopoulos, M., & Bennett, P. (1999). Word learning in the context of referential and salience cues. *Developmental Psychology, 35*(1), 60–68. https://mu.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=1998-03083-005&site=ehost-live&scope=site

Poole, K. L., & Schmidt, L. A. (2021). Vigilant or avoidant? children's temperamental shyness, patterns of gaze, and physiology during social threat. *Developmental Science, 24*(6), e13118. https://doi.org/https://doi.org/10.1111/desc.13118

Putnam, S. P., Gartstein, M. A., & Rothbart, M. K. (2006). Measurement of fine-grained aspects of toddler temperament: The early childhood behavior questionnaire. *Infant Behavior and Development, 29*(3), 386–401. https://doi.org/https://doi.org/10.1016/j.infbeh.2006.01.004

Quine, W. V. O. (1960). *Word and object*. MIT Press.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Robb, M. B., Richert, R. A., & Wartella, E. A. (2009). Just a talking book? word learning from watching baby videos. *British Journal of Developmental Psychology, 27*(1), 27–45. https://doi.org/https://doi.org/10.1348/026151008X320156

Rothbart, M. K. (1988). Temperament and the development of inhibited approach. *Child Development*, *59*(5), 1241–1250. Retrieved February 8, 2023, from http://www.jstor.org/stable/1130487

RStudio Team. (2020). *Rstudio: Integrated development environment for r*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Rudasill, K. M., Prokasky, A., Tu, X., Frohn, S., Sirota, K., & Molfese, V. J. (2014). Parent vs. teacher ratings of children's shyness as predictors of language and attention skills. *Learning and Individual Differences, 34,* 57–62. https://doi.org/https://doi.org/10.1016/j.lindif.2014.05.008

Smith Watts, A. K., Patel, D., Corley, R. P., Friedman, N. P., Hewitt, J. K., Robinson, J. L., & Rhee, S. H. (2014). Testing alternative hypotheses regarding the association between behavioral inhibition and language development in toddlerhood. *Child Development*, *85*(4), 1569–1585. https://doi.org/https://doi.org/10.1111/cdev.12219

Spere, K. A., & Evans, M. A. (2009). Shyness as a continuous dimension and emergent literacy in young children: Is there a relation? *Infant and Child Development*, *18*(3), 216–237. https://doi.org/https://doi.org/10.1002/icd.621

Spere, K. A., Schmidt, L. A., Theall-Honey, L. A., & Martin-Chang, S. (2004). Expressive and receptive language skills of temperamentally shy preschoolers. *Infant and Child Development*, *13*(2), 123–133. https://doi.org/https://doi.org/10.1002/icd.345

Twomey, K. E., Ma, L., & Westermann, G. (2018). All the right noises: Background variability helps early word learning. *Cognitive Science*, *42*(S2), 413–438. https://doi.org/https://doi.org/10.1111/cogs.12539

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wieser, M. J., Pauli, P., Alpers, G. W., & Mühlberger, A. (2009). Is eye to eye contact really threatening and avoided in social anxiety?—an eye-tracking and psychophysiology study. *Journal of Anxiety Disorders*, *23*(1), 93–103. https://doi.org/https://doi.org/10.1016/j.janxdis.2008.04.004

Yurovsky, D., & Frank, M. C. (2017). Beyond naïve cue combination: Salience and social cues in early word learning. *Developmental Science*, *20*(2), e12349. https://doi.org/https://doi.org/10.1111/desc.12349

Zhang, L., Eggum-Wilkens, N. D., Eisenberg, N., & Spinrad, T. L. (2017). Children's shyness, peer acceptance, and academic achievement in the early school years. *Merrill-Palmer Quarterly*, *63*(4), 458–484. Retrieved February 16, 2023, from http://www.jstor.org/stable/10.13110/merrpalmquar1982.63.4.0458

Zosh, J. M., Brinster, M., & Halberda, J. (2013). Optimal contrast: Competition between two referents improves word learning. *Applied Developmental Science*, *17*(1), 20–28. https://doi.org/10.1080/10888691.2013.748420

Data, Code and Materials Availability Statement

Raw data, analysis scripts, digital stimuli, and supplementary material used in the current study can be found on the Open Science Framework: https://osf.io/2dhjb. Access to the Early Childhood Behavior Questionnaire (ECBQ) can be requested here: https://research.bowdoin.edu/rothbart-temperament-questionnaires (last retrieved 14/11/2022). The Oxford Communicative Development Inventory (CDI) can be downloaded here: https://www.psy.ox.ac.uk/research/oxford-babylab/research-overview/oxford-cdi (last retrieved 14/11/2022). The Editor agreed an exemption (15/11/2022) to materials-sharing for the ECBQ and Oxford CDI on the basis that both are subject to copyright restrictions.

Ethics statement

Ethical approval for this study was granted by the Lancaster University Research and Ethics Committee. All participants' parents gave informed written consent before taking part in the study.

Authorship and Contributorship Statement

MH wrote the manuscript and collected the data. All authors were involved in study design, analyses, manuscript editing, and contributed intellectually to the manuscript.

Acknowledgements

With thanks to Kirsty Dunn for help with stimuli creation, and the children and caregivers who took part for their time.

Funding Details

This work was supported by the ESRC International Centre for Language and Communicative Development (LuCiD) [ES/L008955/1, ES/S007113/1], an ESRC Future Research Leaders fellowship to KET [ES/N01703X/1] and a British Academy/Leverhulme Trust Senior Research Fellowship to GW [SF150163].

License

Maximizing accuracy of forced alignment for spontaneous child speech

Robert Fromont,
Lynn Clark,
Joshua Wilson Black,
Margaret Blackwood

University of Canterbury, Christchurch, Aotearoa New Zealand

Abstract: Sociophonetic study of large speech corpora generally requires the use of forced alignment — the automatic process of determining the start and end time of each speech sound within the recording — in order to facilitate large-scale automated extraction of acoustic measurements of targeted vowels or consonants. There is an extensive literature evaluating alignment accuracy of a number of forced alignment tools and procedures, processing speech data from a range of languages and dialects. In general, these evaluations use typical adult speech data, often elicited in a controlled laboratory environment. There is little literature on the effectiveness of forced alignment systems on child speech, and none on speech elicited in field environments. This presents a problem for research at the intersection of language acquisition and sociophonetics as there is no established best practice for automatically aligning child speech. Child speech presents special challenges to automated tools, as it includes more variation in speech sounds and voice quality, and non-standard pronunciation and prosody. We evaluated three commonly used forced aligners, the Montreal Forced Aligner (MFA), the Hidden Markov Model Toolkit (HTK) integration provided by the LaBB-CAT corpus analysis tool, and the Penn Aligner (P2FA), using different configurations to force align non-rhotic child speech elicited in a preschool environment. Against many of our expectations, we found that volume of training data trumps similarity to the speech; MFA, using rhotic acoustic models pre-trained on adult speech, performed best. This paper provides a clear methodology for other researchers in sociophonetics to evaluate the success or otherwise of phonetic alignment.

Keywords: child speech; language acquisition; sociophonetics; speech corpora; forced alignment

Corresponding author: Robert Fromont, New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, NZ. Email: robert.fromont@canterbury.ac.nz.

ORCID IDs: https://orcid.org/0000-0001-5271-5487; https://orcid.org/0000-0003-3282-6555; https://orcid.org/0000-0002-8272-5763

Citation: Fromont, R., Clark, L., Wilson Black, J., & Blackwood, M. (2023). Maximizing accuracy of forced alignment for spontaneous child speech. *Language Development Research, 3*(1), 182–210. https://doi.org/10.34842/shrr-sv10

Introduction

There has been a progressive development in the collection and use of large digitised speech corpora containing hundreds of hours of spontaneous speech in sociophonetic research, e.g. the Origins of New Zealand English (ONZE) corpus containing over 3 million words (Gordon et al. (2007)), or the Spoken BNC2014 corpus over 11 million words (Love et al. (2017)). Such corpora are not amenable to painstaking manual alignment to the phone level, which can take 800 times longer than the duration of the speech (Schiel et al. (2012), Section 8.5.1, p. 111, footnote 11). 'Forced alignment', the automated process of locating the start and end times of speech sounds within speech recordings, has been described as 'transformative' by Coto-Solano (2022) (p. 2) allowing the large-scale extraction and study of segments from such corpora.

Although automatically generated alignments of extracted speech sound tokens can be manually checked and adjusted for accuracy, as the number of tokens extracted increases, the practicality of manually checking each and every one decreases. Developing highly accurate tools and procedures for forced alignment is critical, and there is a decades-long literature evaluating different systems and techniques when applied to adult speech. Current best practice in sociophonetics research on adult talkers combines methods which use the most accurate forced alignment configuration, together with procedures for automatically weeding out erroneous tokens after extraction. This method can result in the loss of incredible amounts of data (e.g. Brand et al. (2021) report losing 80% of their data during the filtering process) and yet it still allows measurement and analysis of hundreds of thousands of tokens.[1] Maximising the accuracy of automatic alignment is crucial to minimising such exclusion of data.

Although the literature is well established for typical adult speech, very little work has been done to establish best practices for accurate alignment of child speech. During language development, speech includes more variation in pronunciation (Lee et al. (1999), Assmann & Katz (2000)), duration (Smith (1992), Lee et al. (1999)), and prosody (Athanasopoulou & Vogel (2016)), which can be a challenge for automatic tools that are calibrated for typical adult speech.

After reviewing the current literature on forced alignment of adult and child speech, we describe our own child spontaneous speech corpus, present experiments we ran to determine the most accurate procedure for force aligning our data with three commonly used forced aligners, and the methods we used to measure accuracy. Finally, we present the results of these experiments, and discuss the implications of those results.

[1] In addition to Brand et al. (2021), see recent work by Stuart-Smith et al. (2019). A comprehensive survey of forced alignment used for sociophonetic research is provided by Coto-Solano (2022) (Section 6).

Forced Alignment Tools and Procedures

Since the 1990's a number of computational techniques have been applied to the problem of forced alignment, including Dynamic Time Warping (DTW; Cosi et al. (1991), Coleman (2005)), Hidden Markov Models (HMMs; Young et al. (2006)), and Deep Neural Networks (DNNs; Hawkins et al. (2017)). Forced alignment procedures have sometimes included post-alignment error correction by modelling errors based on a small number of manual alignments (Toledano & Gómez (2002), Adell et al. (2005)).

Most current forced aligners commonly used for phonetics research use one of two HMM-based Automatic Speech Recognition (ASR) software toolkits: the HMM Tool Kit (HTK; Young et al. (2006)) and Kaldi (Povey et al. (2011)).

Although the ASR toolkits themselves support a wide array of options for preparing, processing, and aligning speech data, the forced aligners that have been developed to simplify and automate parts of this process for phonetics generally employ a two-phase process.

Phase one requires three ingredients:

1. a collection of speech recordings,
2. corresponding orthographic transcripts with start and end times of utterances, and
3. a mapping of orthographic spelling to pronunciation using some set of phoneme symbols (usually a pronunciation dictionary).

Hidden Markov Model Gaussian Mixture Models (HMM-GMMs) are trained using the toolkit, which uses Mel Frequency Cepstral Coefficients (MFCC) computed from the audio signal[2], producing a set of acoustic models, either one for each phoneme symbol (monophone models) or one for each distinct cluster of three phonemes (triphone models).

Phase two requires four ingredients:

1. a collection of recordings,
2. corresponding orthographic transcripts,
3. a mapping of orthographic spelling to pronunciation using the same set of phoneme symbols used during phase one, and
4. the acoustic models trained during phase one.

Phase two involves using acoustic models from phase one, either as-is or adapted for each speaker, to align the word pronunciations with the audio, output being a set of

[2]Gaussian Mixture Models (GMMs), are used to model the distribution of the coefficients

start and end times for the words and corresponding phones found in the recordings.

Aligners that use 'pre-trained models' are those where the recordings and transcripts used in phase one are different from those used in phase two. Conversely aligners that use a 'train/align' procedure are those where the same recordings/transcripts are used in *both* phases.

If the recordings in phase one are all from the same speaker, then the models are *speaker-specific*, otherwise they are *speaker-independent*, although some aligners support adapting speaker-independent models to individual speakers during phase two. We refer to the former as *speaker-adapted* models and the latter as *unadapted*.

HTK and Kaldi

HTK and Kaldi are both toolkits for developing ASR systems. They both use HMMs (although Kaldi supports using DNNs instead) and can both be used for training monophone or triphone models.

HTK, developed from 1989 to 2016 by Cambridge University Engineering Department (CUED), is older than Kaldi. Kaldi has been in development since 2009 at Johns Hopkins University, using more 'modern and flexible code' than HTK[3]. While the source code for both toolkits is available, the HTK license requires users to register. Kaldi is released with the Apache License v2.0 licence, and is fully open source.

Current Forced Alignment Systems

Forced alignment systems currently used in sociophonetic research each use their own combination of toolkits, models, and procedures. Widely used systems include:

- Penn Phonetics Lab Forced Aligner (P2FA; Yuan & Liberman (2008)), which uses monophone HTK models pre-trained on American English speech;
- Munich AUtomatic Segmentation (MAUS; Schiel (1999), Schiel (2015)), an HTK-based system with pre-trained models for a wide variety of languages, also available via BAS Web Services[4](Kisler et al. (2017));
- Prosodylab Aligner (Gorman et al. (2011)), an HTK-based system that allows for training of new acoustic models;
- Montreal Forced Aligner (MFA; McAuliffe et al. (2017)), the successor of Prosody-Lab Aligner[5], built on Kaldi's HMM capabilities, including speaker-adapted models, and supporting both pre-trained triphone models (acoustic models and pro-

[3]https://www.kaldi-asr.org/doc/about.html
[4]https://www.bas.uni-muenchen.de/Bas/BasMAUS.html
[5]McAuliffe et al. (2017) p. 1.

nunciation dictionaries for a wide range of languages and varieties are available) and also a train/align mode of operation;
- LaBB-CAT (Fromont & Hay (2012)), a speech corpus management system that integrates with HTK, P2FA, MFA and BAS Web Services, supporting both pre-trained and train/align procedures; and
- Gentle (Hawkins et al. (2017)), which like MFA is built on Kaldi, but unlike MFA, uses DNNs instead of HMMs[6], and supports English only.

Evaluations on Adult Speech

Over the last three decades, the accuracy of many forced alignment tools and configurations has been evaluated using adult speech.

Factors considered in these evaluations include: which tool set is used (Chen et al. (2004), Adell et al. (2005), Niekerk & Barnard (2009), DiCanio et al. (2013), McAuliffe et al. (2017), Meer (2020)), the amount of data used for training (Toledano & Gómez (2002), Chen et al. (2004), Brognaux et al. (2012), Fromont & Watson (2016)), speech style (e.g. read vs. spontaneous) (Chen et al. (2004), Fromont & Watson (2016)), whether monophone or triphone models are used (Toledano & Gómez (2002), Brognaux et al. (2012), McAuliffe et al. (2017)), using pre-trained models or the train/align procedure (Niekerk & Barnard (2009), Brognaux et al. (2012), Fromont & Watson (2016), McAuliffe et al. (2017), Gonzalez, Grama, et al. (2018)), using speaker-independent or speaker-specific models (Toledano & Gómez (2002), Niekerk & Barnard (2009), Brognaux et al. (2012)), how finely chunked the speech is (Chen et al. (2004)), applying automated post-alignment corrections based on a manual aligned sample (Toledano & Gómez (2002), Adell et al. (2005)), or by force-aligning data recursively, adding more data for each new cycle (Moreno et al. (1998), Gonzalez, Grama, et al. (2018)). The literature includes data from different languages (e.g. Afrikaans, English, French, isiZulu, Matukar Panau, Russian, Setswana, Spanish) and language varieties (e.g. American, Australian, Blackburn, Hastings, Liverpool, Manchester, New Zealand, Sunderland, and Westray English), including cases where models were pre-trained on a different language (Niekerk & Barnard (2009), DiCanio et al. (2013), Babinski et al. (2019), Tang & Bennett (2019)) or variety (Fromont & Watson (2016), MacKenzie & Turton (2020)) from the speech being aligned.

Various metrics have been used for comparing manual alignments with automatic ones, including comparing aggregate acoustic measurements (pitch peak, vowel space, and mean duration) resulting from automatic and manual alignments (Babinski et al. (2019)), error thresholds for absolute differences in boundaries (Cosi et al. (1991), Toledano & Gómez (2002), DiCanio et al. (2013), McAuliffe et al. (2017), Tang & Bennett (2019), Meer (2020), Gonzalez, Grama, et al. (2018), Gnevsheva et al. (2020)) or interval

[6]Early versions of MFA included the possibility of using DNNs, but MFA version 2.0 does not

mid-points (Gonzalez, Travis, et al. (2018)), mean / median differences between boundaries (Chen et al. (2004), Gorman et al. (2011), McAuliffe et al. (2017), Gonzalez, Grama, et al. (2018), Tang & Bennett (2019), Meer (2020), Gonzalez et al. (2020)), and the 'Overlap Rate' — the proportional degree of overlap of intervals (Niekerk & Barnard (2009), Fromont & Watson (2016), Gonzalez, Travis, et al. (2018), Gonzalez et al. (2020)).

General conclusions from the literature are that the finer the data is chunked the better and that speaker-specific models are more accurate than speaker-independent models, as are models trained on more data. A mismatch in speech style between the training and alignment data leads to lower accuracy, and using a sample of manual alignments to model post-alignment corrections also boosts accuracy. There is conflicting evidence about whether monophone or triphone models are more accurate. HMM-based systems represent the current state of the art, with a recent preference towards Kaldi-based MFA rather than older HTK-based ones (Gonzalez, Grama, et al. (2018), Gonzalez et al. (2020)).

Evaluations on Child Speech

Work on forced alignment has skewed towards 'high resource' data, i.e. 'mainstream' languages such as English, and high-status varieties of those languages, such as US English. This skew also has a demographic dimension. Development and evaluation of forced alignment tends to use readily available non-pathological adult speech.

However other types of speech also warrant sociophonetic research; child speech has special challenges not usually present in most adult speech. As children are still in the process of developing their language faculties, they show more variability in their phonology, volume, and articulation. The authors have also found unusual prosodic phenomena such as mid-word pauses in our own data (Fromont et al. (2022)).

Alignment accuracy with child speech has only recently received any attention from researchers. Knowles et al. (2018), Mahr et al. (2021), and Szalay et al. (2022) have performed some evaluations which we now describe. Knowles et al. (2018) investigated the effect of various factors on the accuracy of forced alignment of child speech, using a specific forced alignment tool, ProsodyLab-Aligner. They used two corpora of child speech: one comprising 2 hours of spontaneous speech by a single Canadian English speaking child at different ages (1;5 - 3;6), and another including 5 hours of single-word controlled speech by 40 girls and 41 boys aged between two and six years, speaking US English recorded in a laboratory.

Using the attributes of the corpora themselves, they examined the effects of speech style and speaker age. They also compared alignments produced using different types of training data: adult speech only, adult and child speech, and child speech only, training both speaker-independent models and speaker-specific models. In addition they

compared the use of two different dictionaries: a 'standard' dictionary (the CMU Pronouncing Dictionary, Rudnicky & Weide (2014)), and a dictionary manually customised to match the child's speech.

They concluded that controlled speech had more accurate alignment than spontaneous speech,[7] the speech of older children was more accurately aligned, child-only models performed better, and the customised dictionary, which more closely matched the child's actual speech, performed better than a 'standard' dictionary. Vowels and sibilants were best aligned. The best accuracies produced, using their midpoint overlap metric (see below), were 75%-90%.

Mahr et al. (2021) compared different forced aligners - MFA, Kaldi with triphone models, Prosodylab Aligner, and P2FA - using a corpus of 42 US English speaking children aged between 3 and 6 years, recorded in a laboratory. Unlike Knowles et al. (2018), the utterances were generally sentences (up to 60 per participant) rather than single words[8], but were still highly controlled. They found that MFA using models pre-trained on adult speech produced the best alignments, with 86% accuracy (using midpoint overlap). Again, vowels were the best aligned segments.

Szalay et al. (2022) have also evaluated forced aligners on child speech, comparing the MAUS HTK-based aligner with three custom aligners trained using Kaldi's DNN functionality, rather than using HMMs. Their test data were 153 single words elicited from 11 Australian English (AusE) speaking children (7 boys and 4 girls) aged between 4;10 and 11;11. Their custom aligners differed by training data; one was trained on AusE speaking adults, another was trained on speech by similar aged children speaking a different dialect – American English (AmE) – and the third was trained on a mixture of adult AusE and child AmE speech.

They found that the custom aligners trained on adult AusE training data, and the aligner that combined this with AmE child data, had similar high comparative accuracy – with 65% and 66% boundaries within 20ms of the manual boundary, and mean Overlap Rate of 0.74 and 0.73, respectively – better than the aligner that used AmE child speech alone, with 46% accuracy and 0.71 mean Overlap Rate, and MAUS with 59% accuracy and 0.69 mean Overlap Rate. They conclude that matching dialect is more important than matching age.

Our Data

We have a growing corpus of New Zealand children performing an oral language assessment task at their pre-school. Each child heard a story and was asked to re-tell it.

[7] This may have been an caused by the single-word utterances being more finely chunked than the spontaneous utterances

[8] Mahr et al. did not report the total duration of their recordings.

The initial corpus for forced alignment included 38 children (21 boys, 17 girls) aged 3;6 - 4;11.

The literature on adult and child forced alignment would appear to offer clear guidelines for aligning a corpus of child speech.

- The more similar the training and alignment speech, the better; the speaker's own speech is best (i.e. speaker-specific models) but if not, speaker-independent models trained on similar speech work better.
- The closer the dictionary is to the actual pronunciations, the better; a dictionary for the same language variety (with the same phoneme inventory, rhoticity etc.) should be preferred.
- The more training data, the better.

However our initial attempts to force align the speech using LaBB-CAT's default HTK-based training of speaker-specific models and a non-rhotic dictionary suitable for New Zealand English (NZE) produced poor results. We suspected that this kind of corpus falls within a gap in the forced alignment literature.

Although the literature is clear that speaker-specific models are preferable, it is also necessary to have *enough* training data to produce reliable models. Fromont & Watson (2016) found that, for NZE, at least five minutes of speech is required for each speaker for the Overlap Rate to plateau between 0.5 and 0.6[9]. The most verbose child in our corpus spoke for slightly less than three minutes, and many spoke much less than this; the least amount of speech for a single child was sixteen seconds.

We considered using speaker-independent models, either by grouping children in our corpus together in order to train on more than five minutes of speech; our corpus contains 29 minutes child speech, or 46 minutes including adult examiner speech. Or we could use pre-trained models, which are trained on much more data than our corpus contains. However, most models available for English are pre-trained on adult US English speech, which we suspected would be too different from the speech in our corpus.

Almost all of the data used for evaluation in the child speech forced alignment literature was controlled speech; short predictable sentences, and often single words, elicited in a sound-attenuating laboratory environment. But our corpus is spontaneous speech, and is field data recorded in environments with background noise. In many cases the speech is low volume or the child is whispering. The literature appears to have no recommendation for these circumstances; Mahr et al. (2021) are clear about

[9] Overlap Rate is a value between 0 meaning no overlap at all, and 1 meaning perfect overlap; see the section called Overlap Rate for details.

this: "we are hesitant to extrapolate beyond elicited laboratory speech."[10] Furthermore, in some cases the speech is articulated in a manner that's so divergent from adult norms, that even the correct transcription is debatable.

Faced with many doubts about how to proceed, we performed a number of experiments in order to determine 1) which tool/procedure would yield the most accurate alignments, and 2) how the resulting accuracy measured up against accuracies reported in the literature. We expected some configuration involving a non-rhotic dictionary and training on some mix of the children's own speech to result in the most accurate alignments, but that the best accuracy would still be lower than in other studies, due to the age of the speakers and the spontaneous nature of the utterances.

Methods

We compared three commonly used HMM-based aligners, LaBB-CAT's HTK forced-alignment, P2FA (also built on HTK), and MFA (built on Kaldi), and different alignment procedures using those tools:

- train/align with speaker-specific models
- train/align with speaker-independent models
- pre-trained models using a pronunciation dictionary matching our non-rhotic NZE data
- widely-used pre-trained models using a rhotic pronunciation dictionary

In order to easily and reproducibly automate specific configurations, we used LaBB-CAT, which integrates with all three aligners[11], and includes the `nzilbb.labbcat` R package[12], allowing the implementation of an R script to precisely specify forced alignment configurations, and run forced alignment on different subsets of the corpus.

We used ten different forced alignment configurations, which are all easily configurable options with the chosen forced aligners, requiring the minimum manual intervention. The train/align configurations generally use the default options for the given forced aligner (except where otherwise noted), and the pre-trained model configurations use models and dictionaries that are readily available. They represent options that were not only convenient for us to set up quickly for our own LaBB-CAT-based corpus, but also would be easily configured for other sociophonetic research with similar data, either via LaBB-CAT, or in the case of MFA and P2FA, independently of LaBB-CAT

[10] Mahr et al. (2021), p. 2221.

[11] Although LaBB-CAT integrates with BAS Web Services, we could not try MAUS for forced alignment, because our data cannot be shared with a third party

[12] Fromont (2023)

by using the command line interfaces of those forced aligners. The configurations are compared in Table 1. We describe them in detail now.

LaBB-CAT-HTK configurations

The configurations we refer to as 'LaBB-CAT-HTK' use LaBB-CAT's direct integration with the HTK toolkit, which automates the eight steps for training acoustic models with HTK laid out by Young et al. (2006) in Chapter 3 of 'The HTK Book'.

For all train/align configurations using LaBB-CAT-HTK, the same pronunciation dictionary was used: the CELEX English lexicon (Baayen et al. (1995)), a non-rhotic lexicon based on 'British English', supplemented to include words not present in the original lexicon, including non-standard child wordforms such as "comed", "goed", "runned", etc. Phonemic transcriptions are encoded using CELEX's 'DISC' phoneme symbols[13].

The initial base-line configuration was for speaker-specific models; each child's speech was aligned using models trained only on their own speech (*Speaker specific* in Table 1). We also specified three speaker-independent configurations which grouped speakers together for the training phase in groups of increasing size and decreasing speaker similarity. Firstly, speakers were grouped by gender; each child's speech was aligned using models trained on speech of children of the same gender (*Gender specific* in Table 1). Secondly, one set of speaker-independent models were trained using the speech of all children together (*Child independent* in Table 1). Thirdly, one set of speaker-independent models were trained using the speech of all children and also adults in the corpus (*Speaker independent* in Table 1). All speaker-independent models were trained on more than five minutes of speech.

P2FA

The final HTK-based configuration uses the P2FA pre-trained models (*P2FA* in Table 1) in order to compare accuracy of the LaBB-CAT-HTK train/align configurations above with this commonly-used aligner. These models use ARPAbet phoneme symbols[14] that are different from those used by CELEX, and are trained on rhotic US English adult speech[15]. As a result, this configuration used a supplemented version of the CMU Pronouncing Dictionary (CMUdict).[16]

[13] Appendix A includes a table showing how these symbols relate to other symbol sets, and they are described in section 2.4.1 of the CELEX English manual included with Baayen et al. (1995)

[14] See Appendix A.

[15] The P2FA models were trained on 25.5 hours of speech by adult American English speakers, specifically speech of eight Supreme Court Justices selected from oral arguments in the Supreme Court of the United States (SCOTUS) corpus (Yuan & Liberman (2008)).

[16] See Rudnicky & Weide (2014)

MFA configurations

By default MFA uses a train/align procedure that first trains speaker-independent models using all speech, and then adapts these models to each speaker, so that the final alignments use speaker-specific models. Our first MFA configuration used this procedure, using the same CELEX pronunciation dictionary as used by the LaBB-CAT-HTK configurations (*Speaker adapted* in Table 1).

MFA also supports using a variable number of HMM states; each model uses fewer or more states depending on what type of phoneme is being modelled (e.g. fewer states for certain stops, but more for diphthongs). In order to achieve this, MFA requires the phonemic transcriptions to use a specific set of IPA symbols, so we used a supplemented dictionary based on a non-rhotic 'British English' dictionary supplied by MFA[17] (*Variable state* in Table 1).

MFA provides different sets of pre-trained models, so our final three configurations used pre-trained models and corresponding dictionaries. The first two configurations use 'General American English' models using a rhotic dictionary encoded with the same ARPAbet symbols as used by P2FA[18]. The first configuration uses the models 'as-is', without adapting the models to each speaker before alignment (*GAM Unadapted* in Table 1), and the second includes the speaker adaptation step (*GAM Speaker adapted* in Table 1) in order to be able to determine how much difference the speaker adaptation of the models might make with our child speech data. The last configuration uses models trained on different varieties of English using a non-rhotic 'UK English' dictionary encoded using IPA (*UK Speaker adapted* in Table 1)[19]. This final configuration includes much more training data, including non-rhotic (as well as rhotic) varieties of English, and a non-rhotic dictionary, so we suspected it might provide more accurate alignments for our non-rhotic NZE speech than the GAM-based configurations above. Because the dictionary is non-rhotic, as is much of the training data, it is marked as such in Table 1.

[17] See https://mfa-models.readthedocs.io/en/latest/dictionary/English/English%20%28UK%29%20MFA%20dictionary%20v2_0_0a.html.

[18] The English (US) ARPA acoustic model v2.0.0a (McAuliffe & Sonderegger (2022b)) was trained on speech by 2484 American English speakers from the LibriSpeech English corpus (Panayotov et al. (2015)) - for more information see https://mfa-models.readthedocs.io/en/latest/acoustic/English/English%20%28US%29%20ARPA%20acoustic%20model%20v2_0_0a.html

[19] English MFA acoustic model v2.0.0a (McAuliffe & Sonderegger (2022a)) trained on a number of varieties of English from the following corpora: 2479.95 hours from Common Voice English v8.0 (Ardila et al. (2020)), 982.3 hours from Librispeech English (Panayotov et al. (2015)), 124.31 hours from The Corpus of Regional African American Language (Kendall & Farrington (2018)), 5.77 hours from Google Nigerian English (Butryna et al. (2019)), 31.29 hours from the Open-source Multi-speaker Corpora of the English Accents in the British Isles (Demirsahin et al. (2020)), 56.43 hours from The NCHLT speech corpus of the South African languages (Barnard et al. (2014)), and 7.13 hours from the ARU Speech Corpus (University of Liverpool) (Hopkins et al. (2019)) - for more information see https://mfa-models.readthedocs.io/en/latest/acoustic/English/English%20MFA%20acoustic%20model%20v2_0_0a.html

Table 1: Comparison of forced alignment configurations

Aligner	Model	Training	Non-rhotic	Training Data
LaBB-CAT-HTK	Speaker specific	Train/Align	✓	0.3 – 2.9 min
LaBB-CAT-HTK	Gender specific	Train/Align	✓	13.1 – 15 min
LaBB-CAT-HTK	Child independent	Train/Align	✓	29.1 min
LaBB-CAT-HTK	Speaker independent	Train/Align	✓	46.1 min
P2FA	P2FA	Pre-trained	×	25.5 hours
MFA	Speaker adapted	Train/Align	✓	29.1 min
MFA	Variable state	Train/Align	✓	29.1 min
MFA	GAM Unadapted	Pre-trained	×	982.3 hours
MFA	GAM Speaker adapted	Pre-trained	×	982.3 hours
MFA	UK Speaker adapted	Pre-trained	✓	3687.0 hours

Evaluation of Alignments

Manual alignments, for comparison purposes, were provided by one of the authors, a graduate student in linguistics doing research specific to this data, using Praat (Boersma & Weenink (2001)). The best pronunciation was selected from all possibilities in CELEX for each word, using the 'DISC' phoneme symbols. 613 utterances were manually aligned, totalling 28:32 duration, and including 8,514 aligned segments. Manual alignment took approximately 40 hours.

In order to compare each manually aligned phone with its corresponding automatic counterpart, it was necessary to create a mapping between the two sets of alignments. This was complicated by two factors: a) each word may have a different phonemic transcription in the two alignments, because different dictionaries might use different phonemes to transcribe the word,[20] and forced alignment systems can select different pronunciations among all possible pronunciations of a word,[21] b) each dictionary employs a different set of symbols for each phoneme,[22] and don't necessarily use the same phoneme inventories.[23]

[20] e.g. the word "for" is transcribed with two phonemes in CELEX (`f$`), but with three in CMUdict (`F AO1 R`)

[21] e.g. CELEX transcribes the word "and" variously as `{nd` (ænd), `@nd` (ənd), `@n` (ən), `Hd` (n̩d), `H` (n̩), `F` (ŋ̩), or `C` (ŋ̩).

[22] e.g. the word "transcription" is transcribed using the 'DISC' symbols in CELEX, `tr{nskrIpS@n`, the ARPAbet symbols in CMUdict, `T R AE2 N S K R IH1 P SH AH0 N`, and using the IPA in the MFA 'UK English' dictionary, t r æ n s c r ɪ p ʃ ə n.

[23] e.g. the CELEX includes diphthongs `7` (NEAR), `8` (SQUARE) and `9` (CURE), but in CMUdict they are transcribed as multiple phonemes: `IY R`, `EH R`, and `UH R` respectively, and are similarly mismatched in MFA's 'UK English' dictionary, ɪ ə, ɛː, and ʊ ə respectively

In order to ensure the best possible mapping between different alignments, we used a common Minimum Edit Distance algorithm (Wagner & Fischer 1974), modified to ensure matching of similar phonemes across phoneme sets. Appendix A provides a table showing direct correspondences assumed between different symbol sets. The arrows in Figure 1 illustrate how these mappings work; despite the presence of inserted/deleted segments (coloured grey), and also despite the difference in encoding of the segment labels (the manual alignments above use CELEX 'DISC' symbols, while the automatic alignments below use ARPAbet symbols), the algorithm correctly maps corresponding phones to each other.

The literature includes a wide array of metrics for comparing alignments. We wanted to be able to compare our child NZE accuracy with the adult NZE accuracy reported by Fromont & Watson (2016)[24], and that reported by Gonzalez et al. (2020)[25], who reported Overlap Rates of 0.569 and 0.646 respectively. We also wanted to compare accuracies with other evaluations that used laboratory-based child speech; Knowles et al. (2018) reported 75%-90% accuracy using what we call 'Midpoint Containment', and Mahr et al. (2021) reported 86% accuracy using the same metric. In addition Szalay et al. (2022, Table 1.) reported Overlap Rates of 0.69-0.74. We report both of these metrics in our results purely to enable comparison with results from these previous experiments.

Both metrics are independent of the units used, and neither involve arbitrary thresholds to be decided.

Overlap Rate

Paulo & Oliveira (2004) devised Overlap Rate (OvR) as a measure of how much two intervals overlap, independent of their absolute durations. OvR is a value between 0, where the two intervals being compared do not overlap at all, and 1, where the two intervals have the same start and end times. OvR is calculated as follows:

$$OvR = \frac{CommonDur}{DurMax} = \frac{CommonDur}{DurRef + DurAuto - CommonDur},$$

where $CommonDur$ is the duration in common between the automatically aligned and manually aligned segments, $DurRef$ is the duration of the manually aligned segment, and $DurAuto$ is the duration of the automatically aligned segment. $DurMax$ is the maximum duration of the sound file covered by the pair of segments.

Figure 1 visualises how this works; the automatic alignment of the first vowel overlaps with only a third of the corresponding manual alignment, so OvR is 0.333. The second

[24]Fromont & Watson (2016), Section 4.1, p418
[25]Gonzalez et al. (2020) p6, Figure 2.

manually aligned vowel only covers half of the duration of the corresponding automatic alignment, so OvR is 0.5. For the final consonant, both alignments completely overlap each other, resulting in an OvR of 1.

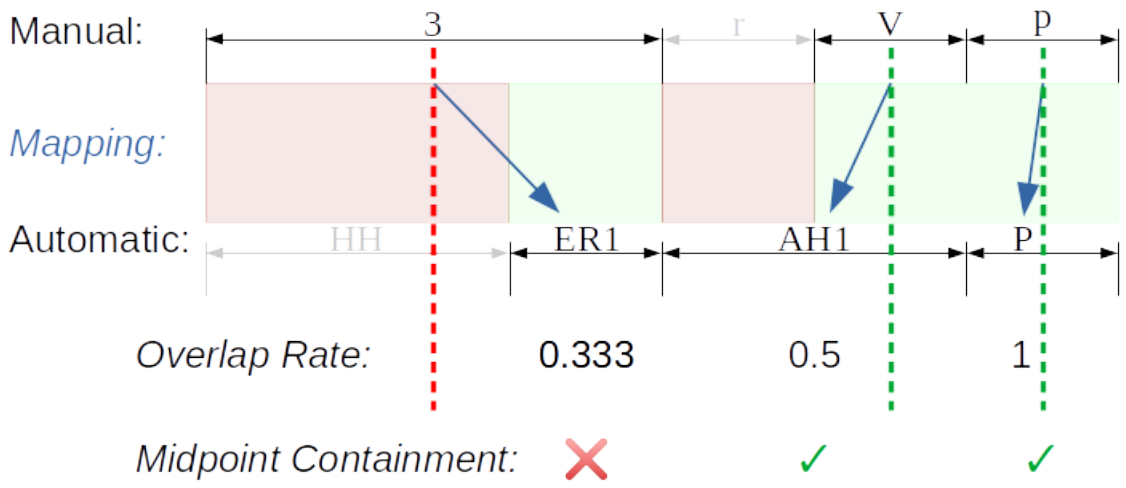


Figure 1. ***Example of mapping manually to automatically aligned phones, and metric computation***

Midpoint Containment

Knowles et al. (2018) devised a measure that calculates the percentage of segments that are 'approximately correct', defined as follows: 'the force-aligned segment overlapped with the midpoint of the corresponding manually aligned phone.'[26] Mahr et al. (2021) use the same metric, calling it a 'gross measure'.[27] Here we prosaically but descriptively call it "midpoint containment".

Figure 1 illustrates how alignments may match or not; the midpoint of the first manually aligned vowel falls outside the bounds of the corresponding automatic interval, so these alignments do not match. For both the overlapping second vowel, and the perfectly aligned final consonant, the manual alignment's midpoint falls within the bounds of its automatic counterpart, so these alignments match.

[26]Knowles et al. (2018) p. 2491, under "Comparisons"
[27]Mahr et al. (2021), p. 4, under "Outcome Variables"

Expectations

Given the general conclusions from the literature our expectations were as follows:

1. Overall performance would be lower than with adult speech, i.e. OvR will be lower than 0.646 (Gonzalez et al. (2020)) and also 0.569 (Fromont & Watson (2016)), because child speech is more varied than adult speech.
2. Overall performance would be lower than with controlled child speech, i.e. Midpoint Containment would be lower than 86% (Mahr et al. (2021)), 75% (Knowles et al. (2018)), and also lower than the 0.69 mean OvR reported by Szalay et al. (2022), because spontaneous speech is more varied than controlled speech.
3. Models trained on child speech would be better than those trained on adult speech, because in general the more similar the training and alignment speech, the better.
4. Non-rhotic dictionaries/models should perform better than rhotic ones; rhotic alignments will include alignments for post vocalic /ɹ/ phones that are not present in our non-rhotic NZE speech, so neighbouring automatic phones will overlap less with their manual counterparts.
5. MFA will perform better than the HTK-based aligners (LaBB-CAT-HTK and P2FA in our case), as found by González et al. (Gonzalez, Grama, et al. (2018), Gonzalez et al. (2020)).
6. Vowels will be the best aligned segments, as previously reported by Knowles et al. (2018) and Mahr et al. (2021).

Results

Table 2 compares both Overlap Rate and Midpoint Containment percentages for each of the forced alignment configurations. All train/align configurations have a mean OvR less than 0.3, and less than 50% Midpoint Containment, with the MFA configurations performing worse than the LaBB-CAT-HTK ones. Conversely, all configurations using models pre-trained on adult speech have a mean OvR greater than 0.3; the P2FA models produce a mean OvR of 0.345, the MFA GAM Unadapted models, 0.429, the MFA UK Speaker adapted models, 0.440, and the MFA GAM Speaker adapted models, the highest mean OvR at 0.458. In terms of Midpoint Containment, 48% of the P2FA alignments contain the midpoint of the corresponding manual alignment, and more than 50% of MFA alignments contain the manual alignment midpoint; 59% for GAM Unadapted models, 62% for UK Speaker adapted models, and 63% for GAM Speaker adapted models.

Figure 2 shows the distributions of Overlap Rates for each configuration. All train/align configurations have a third quartile of less than 0.6, and a first quartile of 0 (along with the P2FA pre-trained models). The *variable state* train/align models perform worst of

Table 2: Mean OvR and percent midpoint-contained, for each forced alignment configuration, with the best performing configuration in bold typeface

Aligner	Model	Training	Non-rhotic	Mean OvR	%
LaBB-CAT-HTK	Speaker specific	Train/Align	✓	0.228	37
LaBB-CAT-HTK	Gender specific	Train/Align	✓	0.261	42
LaBB-CAT-HTK	Child independent	Train/Align	✓	0.298	46
LaBB-CAT-HTK	Speaker independent	Train/Align	✓	0.276	42
P2FA	P2FA	Pre-trained	×	0.345	48
MFA	Speaker adapted	Train/Align	✓	0.239	34
MFA	Variable state	Train/Align	✓	0.155	22
MFA	GAM Unadapted	Pre-trained	×	0.429	59
MFA	**GAM Speaker adapted**	**Pre-trained**	×	**0.458**	**63**
MFA	UK Speaker adapted	Pre-trained	✓	0.440	62

all, with a median of 0, although curiously there are a number of outliers with OvR greater than 0.5. Only the pre-trained MFA models manage a first quartile greater than 0, and all have a third quartile greater than 0.7.

Figure 3 shows the distributions of Overlap Rates for each configuration, broken down by segment category. For the HTK-based tools (the left five configurations), there appears to be little differentiation in accuracy between different segment types. But for MFA (the right five configurations), vowels in particular seem to be very inaccurate for train/align configurations, but quite accurate for pre-trained configurations. Apart from those using MFA pre-trained models, none of the configurations had a first quartile higher than zero for any segment category.

Figure 3 also shows that, although the *GAM Speaker adapted* and *UK Speaker adapted* configurations have similar first and third quartiles for fricatives, the second quartile for *GAM Speaker adapted* is somewhat lower than for *UK Speaker adapted*. The mean fricative OvR for *GAM Speaker adapted* is 0.359 and the corresponding mean for *UK Speaker adapted* is 0.411.

Discussion

The most obvious result is that expectation 3., that 'models trained on child speech would be better than those trained on adult speech', was not borne out by the configurations we tested. All configurations that used only adult data were more accurate than all configurations that used any child data. This surprised us and apparently contradicts Knowles et al. (2018): 'For both corpora, training on adult speech led to poorer

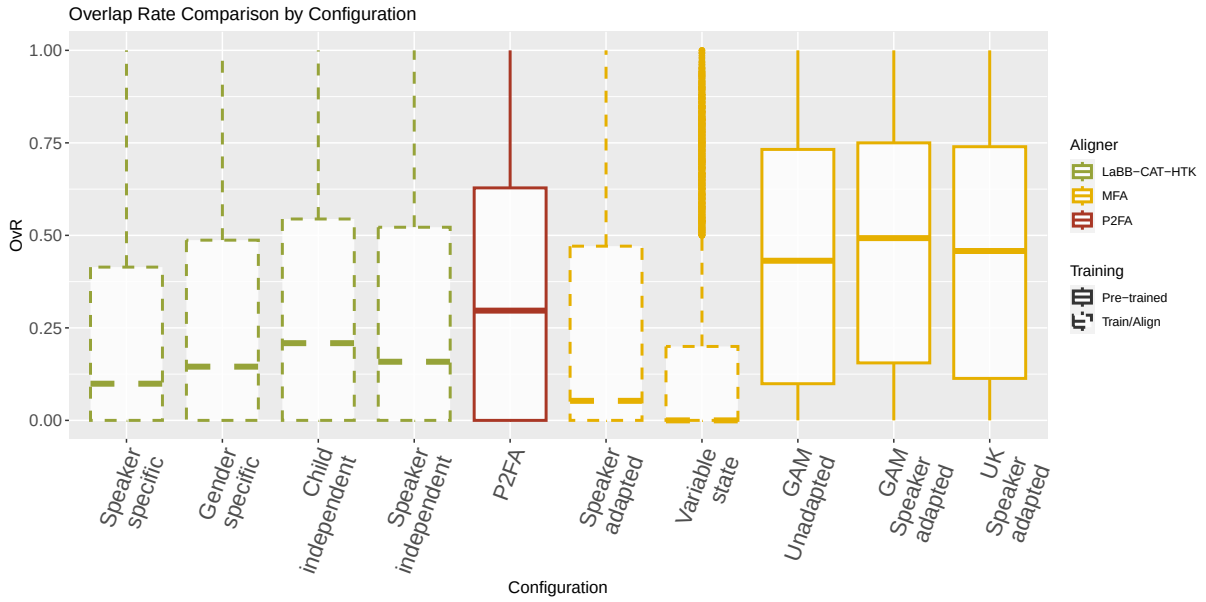


Figure 2. ***Overlap Rate distributions by Configuration. Green lines indicate LaBB-CAT-HTK configurations, dark red lines indicate the P2FA configuration, and yellow lines indicate MFA configurations. Filled lines indicate pretrained configurations, while dashed lines indicate Train/Align configurations.***

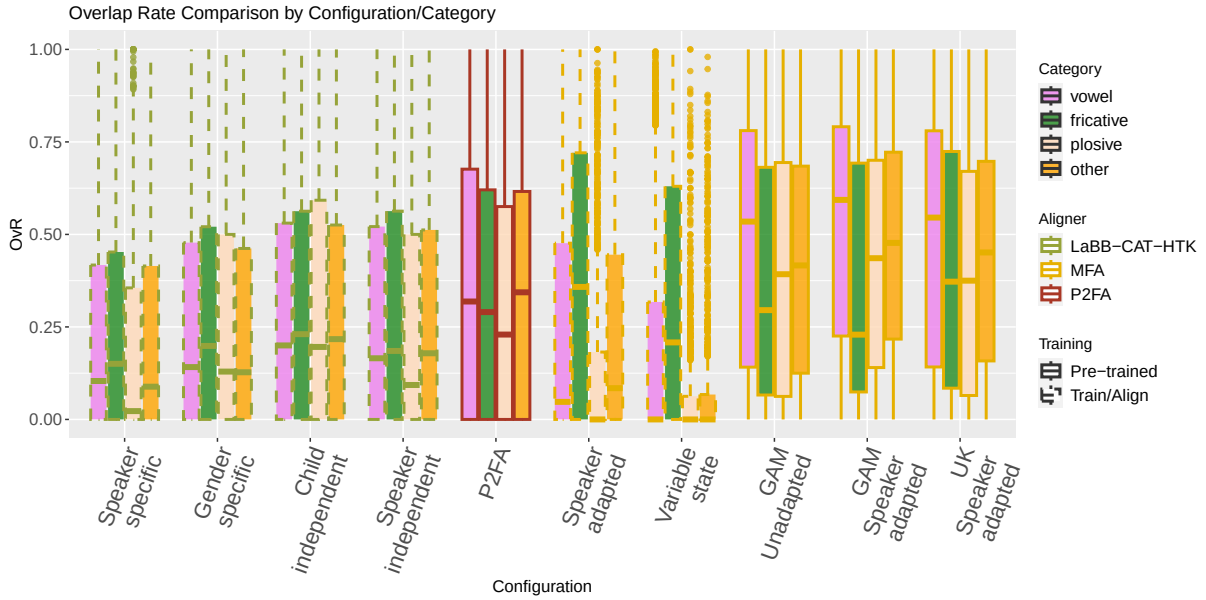


Figure 3. ***Overlap Rate distributions by Configuration, broken down by Segment Category. Line colour indicates the aligner used. Box colour indicates segment category.***

accuracy than training on child speech and can be summarized as follows: Adult-only < Adult–child < Child speech only'[28] although it's in line with Szalay et al. (2022) (section 4.1, p. 39) for whom the best aligners included adult data.

There are various possible explanations for this. Adult speech should be less phonologically varied than child speech, and represents the 'target' forms that children have not yet settled on; perhaps this stability leads to more discerning acoustic models. Or perhaps it's simply because there was more adult speech (25 - 3687 hours) than child speech (29.1 minutes) to train on. Knowles et al. had ten times this amount of child data (5 hours), which yielded alignments that were more accurate than models trained on adult speech (10 hours), so this may indicate that the latter explanation is correct: volume of training data trumps similarity to the speech to be aligned. It's clear that in some cases adult training data leads to higher accuracy for child speech, but further work is required to settle the question of whether this is because of the magnitude of the training data or its qualities.

Another surprise is that the configurations using a rhotic dictionary outperformed those using a non-rhotic dictionary. Using a rhotic dictionary for non-rhotic spontaneous speech inserts tokens of post-vocalic /ɹ/ which do not correspond to the speech. This inevitably decreases alignment accuracy[29], as the extra phone will invade the durations of surrounding phones. This can be seen in Figure 4., which shows an utterance from our corpus, with the correct manual alignment shown above, and the the automatic alignment produced by MFA below. The fourth word, "for", is correctly transcribed with two phonemes, `f $`, but MFA has used the three-phoneme transcription from its rhotic dictionary, `F AO1 R`, the last phone of which is an incorrect insertion taking up most of the duration of the vowel, which has a resulting low OvR of 0.099.

The MFA rhotic dictionary produced marginally better alignments (0.458 mean OvR) than the non-rhotic one (0.440 mean OvR) despite this 'inserted /ɹ/ penalty'. We investigated the incidence of spurious /ɹ/ phones in these alignments, and found that there were only 65 inserted /ɹ/ phones with a mean duration of 74ms, less than 1% of all the phones found in this alignment.[30]

The *English (US) ARPA* models are seemingly so much better than the *English MFA* models that the effect of having extra post-vocalic /ɹ/ tokens is rendered irrelevant. This

[28] Knowles et al. (2018) p. 2492.

[29] If the spurious phones are of zero length, accuracy would not be affected, but there were no zero duration insertions of this type in our data.

[30] Indeed /ɹ/ wasn't even the most common spurious phone; there were more spurious /d/ and /ə/ phones (118 and 68 tokens respectively), mainly representing the final phoneme of the words "and", "the" and "to".

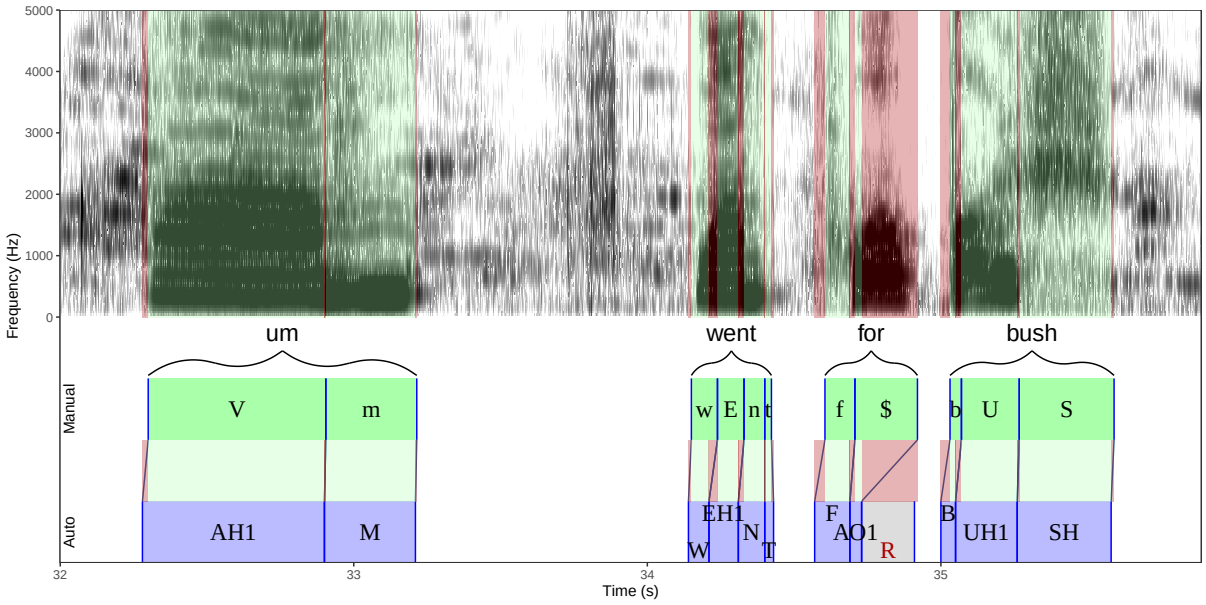


Figure 4. ***Example utterance alignment including inserted rhotic /ɹ/ in the word 'for' - correctly aligned regions are shown in light green, incorrectly aligned regions are shaded in red, and the utterance spectrogram is shown above for reference***

supports and perhaps explains conclusions of Gonzalez et al. (2020)[31] and MacKenzie & Turton (2020)[32] that dictionary/variety don't greatly impact measured performance: the incidence of features that define differences between varieties of the same language (in terms of insertion or deletion of segments) are not frequent enough to make much difference to overall alignment accuracy. However, this contradicts the advice of Szalay et al. (2022) (section 4.1, p. 39): "using a dialect matched, AusE pronunciation dictionary is recommended", and the impact of these discrepancies may indeed be important for downstream research that uses the resulting automatic alignments. For example if sociophonetic research is later conducted on word-final vowels, or on rhoticity itself, the spurious /ɹ/ tokens may significantly interfere with the results.

This better performance cannot be explained by differences in training set size; the models used with the GAM English dictionary were trained on under a thousand hours of speech, but still produced better alignments than the models used with the UK English dictionary, which were trained on over three thousand hours of speech. Apart from the amount of training data and the pronunciations, there are two other differences between these configurations: the latter was trained on numerous varieties of English, and the phoneme sets were differently distributed; The GAM English dic-

[31]Gonzalez et al. (2020) p. 9, section 5.
[32]MacKenzie & Turton (2020) p. 11 section 6.

tionary includes 39 stress-marked vowels and 24 consonants encoded in ARPABET[33], where the UK English dictionary includes 22 vowels and 46 consonants including vocalic and aspirated variants encoded with IPA symbols[34]. Investigating the impact each of these factors has is outside the scope of the current experiment, but it's clear from our results that more training data does not inevitably result in better alignments.

When comparing the performance of HTK-based aligners with MFA, the results are also nuanced. MFA did indeed perform better than HTK-based aligners under the conditions we tested, and as the *GAM Unadapted* accuracy is only slightly below *GAM Speaker adapted*, the difference in accuracy apparently doesn't come down to MFA's speaker-adaptation process. But MFA was more accurate only using pre-trained models. MFA produced the worst alignments among the configurations we tested when using a train/align procedure. The amount of training data is probably the important factor here. Michael McAuliffe, the primary software developer of MFA, notes that 3-5 hours of speech is required for good alignments.[35] It seems that under the conditions of our experiment, LaBB-CAT-HTK works better with scarce data than MFA does. More rigorous comparison between these ASR toolkits may well identify forced alignment methods, or attributes of training data, that yield different results. However, under the conditions we were working with – a relatively small amount of child speech, using the default procedures for LABB-CAT-HTK and MFA – train/align forced alignment was more accurate using LaBB-CAT-HTK, although accuracy was low for both aligners.

When compared with results from other studies, our expectations were borne out. Accuracy was lower than with adult speech, as the best mean OvR of 0.458 was lower than both 0.646 (Gonzalez et al. (2020)) and 0.569 (Fromont & Watson (2016)). Similarly, accuracy was lower with our spontaneous speech than with controlled child speech; our best Midpoint Containment of 63% was lower than 86% (Mahr et al. (2021)) and 75% (Knowles et al. (2018)), and our best mean OvR was lower than 0.69 (Szalay et al. (2022)).

Using MFA pre-trained models, vowels were indeed the best-aligned segments, confirming results from Knowles et al. (2018) and Mahr et al. (2021). However, this was not the case with other configurations. Although the *English (US) ARPA* models are marginally better than the *English MFA* models overall, the latter was better at aligning fricatives. As noted earlier with reference to rhoticity, which models/dictionaries turn out to be best depends somewhat on what types of segment will be analysed downstream.

[33]The GAM English phoneme set is shown in Appendix A

[34]The consonant variants of the UK English phoneme set is shown in Appendix A, Table 4

[35]Michael McAuliffe, "How much data do you need for a good MFA alignment?" (24 August 2021): https://memcauliffe.com/how-much-data-do-you-need-for-a-good-mfa-alignment.html

Conclusion

While there is an established literature on forced alignment methodology for adult speech, accuracy with child speech has only recently received any attention from researchers, and the best approach for dealing with field recordings of children has not been established.

We found that MFA, using acoustic models pre-trained on 'General American English', produced the most accurate alignments of spontaneous NZE child speech in our corpus. These alignments were less accurate than is possible with adult speech of the same variety, and with controlled child speech, and all future alignments in our growing corpus will require manual checking/correction.

Although the results of our experiments resolved a practical problem for us, identifying a clear way forward for the force-alignment of our own corpus, we recognise that they are specific to our speech data and the configurations we tried, using conveniently configurable tools designed specifically for sociophonetic research. More rigorous further work would be required to tease apart the relative importance of the various factors – toolkit, technology, data preparation, amount of and nature of the speech, age and dialect of speakers in the training vs. alignment data, etc.

For example somewhere between the half hour of speech we had available for training, and the two to five hours of speech used by Knowles et al. there may be a threshold where training on child speech alone yields better alignments than those produced by using models pre-trained on adult speech. Furthermore, the recursive method of forced alignment studied by Gonzalez, Travis, et al. (2018) may provide a boost in performance. These are questions to be resolved by future investigation, on a larger corpus of child speech.

In addition the present results compared only HMM-based forced alignment. However, Kaldi also supports the use of DNNs for forced alignment. It would be useful to compare performance of DNN-based alignments with HMM-based ones, using Gentle out of the box, or by training custom aligners as done by Szalay et al. (2022). They point out that their best custom aligner was trained on the same AusE dataset as the HMM-based MAUS aligner, which performed the worst, concluding that the difference in performance can be attributed to using Kaldi and DNNs, rather than HTK and HMMs (Szalay et al. (2022), p. 39, section 4.2). If Kaldi alone were used to discover whether DNNs or HMMs produce more accurate alignments, it could be determined whether it's the toolkit or the technology that makes the difference.

We conclude that alignment procedures that work well with adult data are not guaranteed to produce the best results for children. To maximise accuracy, automated alignment of language acquisition corpora requires special attention, and evaluating differ-

ent options on specific corpora is well worth the effort. Even so, our finding with NZE child speech was the same as that of Szalay et al. (2022, p.38 section 4) with AusE child speech: manual correction is still required. We echo MacKenzie & Turton (2020)[36] who recommend that "these aligners are used in the manner for which they were designed — as tools, and not as the complete replacement of a dedicated researcher".

References

Adell, J., Bonafonte, A., Gómez, J. A., & Castro, M. J. (2005). Comparative study of automatic phone segmentation methods for TTS. *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, *1*, I/309–I/312 Vol. 1.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4218–4222. https://aclanthology.org/2020.lrec-1.520

Assmann, P. F., & Katz, W. F. (2000). Time-varying spectral change in the vowels of children and adults. *The Journal of the Acoustical Society of America*, *108*(4), 1856–1866. https://doi.org/10.1121/1.1289363

Athanasopoulou, A. A., & Vogel, I. (2016). Acquisition of prosody: The role of variability. *Speech Prosody*, 716–720.

Baayen, H., Piepenbrock, R., & Rijn, H. V. (1995). *The CELEX Lexical Database (Release 2, CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania. https://catalog.ldc.upenn.edu/LDC96L14

Babinski, S., Dockum, R., Craft, J. H., Fergus, A., Goldenberg, D., & Bowern, C. (2019). A Robin Hood approach to forced alignment: English-trained algorithms and their use on Australian languages. *Proceedings of the Linguistic Society of America*, *4*(1), 3-1-12. https://doi.org/10.3765/plsa.v4i1.4468

Barnard, E., Davel, M., Heerden, C. van, Wet, F., & Badenhorst, J. (2014). The NCHLT Speech Corpus of the South African languages. *SLTU 2014*, 194–200.

Boersma, P., & Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

Brand, J., Hay, J., Clark, L., Watson, K., & Sóskuthy, M. (2021). Systematic co-variation of monophthongs across speakers of New Zealand English. *Journal of Phonetics*, *88*, 101096. https://doi.org/https://doi.org/10.1016/j.wocn.2021.101096

Brognaux, S., Roekhaut, S., Drugman, T., & Beaufort, R. (2012). Automatic Phone Alignment - A Comparison between Speaker-Independent Models and Models Trained on the Corpus to Align. *JapTAL*.

Butryna, A., Chu, S. H. C., Demirsahin, I., Gutkin, A., Ha, L., He, F., Jansche, M., Johny, C. C., Katanova, A., Kjartansson, O., Li, C. F., Merkulova, T., Oo, Y. M., Pipatsrisawat, K., Rivera, C. E., Sarin, S., Silva, P. D., Sodimana, K., Sproat, R., … Wibawa,

[36]MacKenzie & Turton (2020) p. 12 section 6

J. A. E. (2019). Google Crowdsourced Speech Corpora and Related Open-Source Resources for Low-Resource Languages and Dialects: An Overview. *2019 UNESCO International Conference Language Technologies for All (LT4All): Enabling Linguistic Diversity and Multilingualism Worldwide*, 91–94. https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.23.pdf

Chen, L., Liu, Y., Harper, M. P., Maia, E., & McRoy, S. (2004). Evaluating Factors Impacting the Accuracy of Forced Alignments in a Multimodal Corpus. *LREC*.

Coleman, J. (2005). *Introducing speech and Language Processing*. Cambridge University Press.

Cosi, P., Falavigna, D., & Omologo, M. (1991). A preliminary statistical evaluation of manual and automatic segmentation discrepancies. *2nd European Conference on Speech Communication and Technology (Eurospeech 1991)*.

Coto-Solano, R. (2022). Computational sociophonetics using automatic speech recognition. *Language and Linguistics Compass*, *16*(9), e12474. https://doi.org/https://doi.org/10.1111/lnc3.12474

Demirsahin, I., Kjartansson, O., Gutkin, A., & Rivera, C. (2020). Open-source Multispeaker Corpora of the English Accents in the British Isles. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6532–6541. https://aclanthology.org/2020.lrec-1.804

DiCanio, C., Nam, H., Whalen, D. H., Timothy Bunnell, H., Amith, J. D., & García, R. C. (2013). Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America*, *134*(3), 2235–2246. https://doi.org/10.1121/1.4816491

Fromont, R. (2023). *nzilbb.labbcat R package* (Version 1.2-0). https://github.com/nzilbb/labbcat-R/

Fromont, R., Black, J. W., Clark, L., & Blackwood, M. (2022). Forced alignment of child speech: Comparing HTK and kaldi-based aligners. *Third Workshop on Sociophonetic Variability in the English Varieties of Australia*.

Fromont, R., & Hay, J. (2012). LaBB-CAT: an Annotation Store. *Proceedings of Australasian Language Technology Association Workshop*, 113–117.

Fromont, R., & Watson, K. (2016). Factors influencing automatic segmental alignment of sociophonetic corpora. *Corpora*, *11*(3), 401–431.

Gnevsheva, K., Gonzalez, S., & Fromont, R. (2020). Australian English Bilingual Corpus: Automatic forced-alignment accuracy in Russian and English. *Australian Journal of Linguistics*, *40*(2), 182–193. https://doi.org/10.1080/07268602.2020.1737507

Gonzalez, S., Grama, J., & Travis, C. (2018). *FoACL: Forced-Alignment Comparison for Linguistics*. https://cloudstor.aarnet.edu.au/plus/s/gyC6vuX5uvc5soG

Gonzalez, S., Grama, J., & Travis, C. E. (2020). Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard*, *6*(1), 20190058. https://doi.org/doi:10.1515/lingvan-2019-0058

Gonzalez, S., Travis, C., Grama, J., Barth, D., & Ananthanarayan, S. (2018). *Recursive forced alignment: A test on a minority language* (pp. 145–148).

Gordon, E., Maclagan, M., & Hay, J. (2007). The ONZE corpus. In J. C. Beal, K. P. Cor-

rigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora: Volume 2: Diachronic databases* (pp. 82–104). Palgrave Macmillan UK. https://doi.org/10.1057/9780230223202_4

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics, 39*, 192–193.

Hawkins, M., Strob, R. B., & andrakeshshrestha31, D. B. B. (2017). *Gentle*. http://lowerquality.com/gentle/

Hopkins, C., Graetzer, S., & Seiffert, G. (2019). *ARU speech corpus (University of Liverpool)*.

Kendall, T., & Farrington, C. (2018). The Corpus of Regional African American Language. *Version, 6*, 1.

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language, 45*, 326–347. https://doi.org/https://doi.org/10.1016/j.csl.2017.01.005

Knowles, T., Clayards, M., & Sonderegger, M. (2018). Examining Factors Influencing the Viability of Automatic Acoustic Analysis of Child Speech. *Journal of Speech, Language, and Hearing Research, 61*(10), 2487–2501. https://doi.org/10.1044/2018/_JSLHR-S-17-0275

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America, 105*(3), 1455–1468. https://doi.org/10.1121/1.426686

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations [Journal Article]. *International Journal of Corpus Linguistics, 22*(3), 319–344. https://doi.org/https://doi.org/10.1075/ijcl.22.3.02lov

MacKenzie, L., & Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard, 6*(s1), 20180061. https://doi.org/doi:10.1515/lingvan-2018-0061

Mahr, T. J., Berisha, V., Kawabata, K., Liss, J., & Hustad, K. C. (2021). Performance of Forced-Alignment Algorithms on Children's Speech. *Journal of Speech, Language, and Hearing Research, 64*(6S), 2213–2222. https://doi.org/10.1044/2020/_JSLHR-20-00268

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. *Proc. Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

McAuliffe, M., & Sonderegger, M. (2022a). *English MFA acoustic model v2.0.0a*.

McAuliffe, M., & Sonderegger, M. (2022b). *English (US) ARPA acoustic model v2.0.0a*.

Meer, P. (2020). Automatic alignment for New Englishes: Applying state-of-the-art aligners to Trinidadian English. *The Journal of the Acoustical Society of America, 147*(4), 2283–2294. https://doi.org/10.1121/10.0001069

Moreno, P. J., Joerg, C., Thong, J.-M. V., & Glickman, O. (1998). A recursive algorithm for the forced alignment of very long audio segments. *Proc. 5th International Conference*

on Spoken Language Processing (ICSLP 1998), paper 0068. https://doi.org/10.21437/ICSLP.1998-603

Niekerk, D. van, & Barnard, E. (2009). Phonetic alignment for speech synthesis in under-resourced languages. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 880–883. https://doi.org/10.21437/Interspeech.2009-266

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210.

Paulo, S., & Oliveira, L. C. (2004). Automatic Phonetic Alignment and Its Confidence Measures. In J. L. Vicedo, P. Martínez-Barco, R. Muńoz, & M. Saiz Noeda (Eds.), *Advances in natural language processing* (pp. 36–44). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30228-5_4

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N. K., Hannemann, M., Motlícek, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., & Veselý, K. (2011). *The Kaldi Speech Recognition Toolkit*.

Rudnicky, A., & Weide, R. (2014). *Carnegie Mellon University Pronouncing Dictionary*. Carnegie Mellon University; http://www.speech.cs.cmu.edu/cgi-bin/cmudict. http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/

Schiel, F. (1999). *Automatic Phonetic Transcription of Non-Prompted Speech*.

Schiel, F. (2015). A statistical model for predicting pronunciation. *International Congress of Phonetic Sciences*.

Schiel, F., Draxler, C., Baumann, A., Ellbogen, T., & Steffen, A. (2012). *The Production of Speech Corpora*. http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-13693-2

Smith, B. L. (1992). Relationships between Duration and Temporal Variability in Children's Speech. *The Journal of the Acoustical Society of America*, *91*(4 Pt 1), 2165–2174. https://doi.org/10.1121/1.403675

Stuart-Smith, J., Sonderegger, M., Macdonald, R., Mielke, J., McAuliffe, M., & Thomas, E. R. (2019). Large-scale acoustic analysis of dialectal and social factors in English /s/-retraction. *Proceedings of the 19th International Congress of Phonetic Sciences*. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_1322.pdf

Szalay, T., Shahin, M., Ballard, K., & Ahmed, B. (2022). Training forced aligners on (mis)matched data: The effect of dialect and age. *Proceedings of the Eighteenth Australasian International Conference on Speech Science and Technology*, 36–40.

Tang, K., & Bennett, R. (2019). Unite and conquer: Bootstrapping forced alignment tools for closely-related minority languages (Mayan). *Proceedings of the 19th International Congress of Phonetic Sciences*. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_1768.pdf

Toledano, D. T., & Gómez, L. A. H. (2002, May). HMMs for Automatic Phonetic Segmentation. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. http://www.lrec-conf.org/proceedings/lrec2002/pdf/142.

pdf
Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchovaltchev, & Woodland, P. (2006). *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department; http://htk.eng.cam.ac.uk/docs/docs.shtml.
Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics '08*, 5687–5690.

Data, Code and Materials Availability Statement

The script used for configuring forced alignment conditions and processing data, and also the edit paths mapping manually aligned phones to their automatically aligned counterparts, are available using the following URL:

https://doi.org/10.17605/OSF.IO/8R9FP

Ethics statement

Approval for use of this data was gained from the University of Canterbury Human Ethics Committee (Ref 2020/10/ERHEC).

Authorship and Contributorship Statement

- Robert Fromont: Conceptualization, Methodology, Software, Visualization, Writing, Review & editing
- Lynn Clark: Conceptualization, Funding acquisition, Review & editing
- Joshua Wilson Black: Data curation, Visualization, Review & editing
- Margaret Blackwood: Data curation, Review & editing

Acknowledgements

We gratefully acknowledge the support of the Marsden Fund | Te Pūtea Rangahau a Marsden (Application number: 20-UOC-064).

We would like to thank Tristan Mahr for his valuable feedback on an earlier draft, and our anonymous reviewers for taking the time to thoroughly critique our manuscript, whose comments were invaluable for sharpening our thinking and refining our conclusions.

License

Appendix A - Phoneme Symbol Sets

Different English pronunciation dictionaries use different sets of symbols. In many cases, there are quirks that relate to the provenance and purpose of the dictionary; for example the CMU Pronouncing dictionary (CMU Dict) has no symbol for schwa, because unstressed vowels are instead suffixed with 0. Some of the MFA dictionaries use IPA symbols, but transcribe diphthongs in unfamiliar ways, perhaps because of efforts by the developer to develop multi-lingual models. CELEX's 'DISC' symbols are similar to the SAMPA symbols familiar to many linguists, except they conform to the principle that each phoneme can be represented by exactly one character. Below is a table showing how different symbols sets relate to each other.

Table 3: Vowels

Example	IPA	MFA	DISC	CMU Dict[37]
k**i**t	ɪ	ɪ	I	IH
dr**e**ss	ɛ	ɛ	E	EH
tr**a**p	æ	æ	{	AE
str**u**t	ʌ	ɐ	V	AH
f**oo**t	ʊ	ʊ	U	UH
another	ə	ə	@	
fl**ee**ce	iː	iː/**i**	i	IY
b**a**th	ɑː	ɑː	#	AA
l**o**t	ɒ	ɒ	Q	AO
th**ou**ght	ɔː	ɒː	$	AO
g**oo**se	uː	ʉː/ʉ	u	UW
n**ur**se	ɜː	ɜː/ɜ	3	ER
f**a**ce	eɪ	ej	1	EY
pr**i**ce	aɪ	aj	2	AY
ch**oi**ce	ɔɪ	ɔj	4	OY
g**oa**t	əʊ	əw	5	OW
m**ou**th	aʊ	aw	6	AW
n**ear**	ɪə	ɪ ə	7	IY R
sq**uare**	ɛə	ɛː	8	EH R
c**ure**	ʊə	ʊ ə	9	UH R
t**i**mbre	æ̃		c	
dét**en**te	ɑ̃ː	ɑ	q	
l**in**gerie	æ̃ː		0	
bouill**on**	ɒ̃ː		~	

[37] All vowels in CMU Dict's ARPABET encoding have three variants, each suffixed with a digit: 0 for unstressed, 1 for primary stress, and 2 for secondary stress

Table 4: Consonants

Example	IPA	MFA	DISC	CMU Dict
pat	p	p/pʰ/pʲ	p	P
bad	b	b/bʲ	b	B
tack	t	t/tʰ/tʲ	t	T
dad	d	d/dʲ	d	D
cad	k	k/kʰ/c/cʰ	k	K
game	g	g/ɟ	g	G
ba**ng**	ŋ	ŋ	N	NG
mad	m	m/mʲ/ɱ	m	M
nat	n	n/ɲ	n	N
lad	l	l/ɬ/ʎ	l	L
rat	ɹ	ɹ	r	R
fat	f	f/fʲ	f	F
vat	v	v/vʲ	v	V
thin	θ	θ	T	TH
then	ð	ð	D	DH
sap	s	s	s	S
zap	z	z	z	Z
sheep	ʃ	ʃ	S	SH
mea**s**ure	ʒ	ʒ	Z	ZH
yank	j	j	j	Y
had	h	h/ç	h	HH
wet	w	w	w	W
cheap	tʃ	tʃ	J	CH
jeep	dʒ	dʒ	_	JH
lo**ch**	x		x	
baco**n**	ŋ̍		C	
idealis**m**	m̩	m̩	F	
burde**n**	n̩	n̩	H	
dang**l**e	l̩	ɫ̩	P	
ca**r** alarm	*		R	
uh-oh	ʔ	ʔ		

An automated classifier for periods of sleep and target-child-directed speech from LENA recordings

Janet Y. Bang
San José State University, USA

George Kachergis
Stanford University, USA

Adriana Weisleder
Northwestern University, USA

Virginia A. Marchman
Stanford University, USA

Abstract: Some theories of language development propose that children learn more effectively when exposed to speech that is directed to them (target child directed speech, tCDS) than when exposed to speech that is directed to others (other-directed speech, ODS). During naturalistic daylong recordings, it is useful to identify periods of tCDS and ODS, as well as periods when the child is awake and able to make use of that speech. To do so, researchers typically rely on the laborious work of human listeners who consider numerous features when making judgments. In this paper, we detail our efforts to automate these processes. We analyzed over 1,000 hours of audio from daylong recordings of 153 English- and Spanish-speaking families in the U.S. with 17- to 28-month-old children that had been previously coded by human listeners for periods of sleep, tCDS, and ODS. We first explored patterns of features that characterized periods of sleep, tCDS, and ODS. Then, we evaluated two classifiers that were trained using automated measures generated from LENA™, including frequency (AWC, CTC, CVC) and duration (meaningful speech, distant speech, TV, noise, silence) measures. Results revealed high sensitivity and specificity in our sleep classifier, and moderate sensitivity and specificity in our tCDS/ODS classifier. Moreover, model-derived predictions replicated previously-published findings showing significant and positive links between tCDS, but not ODS, and children's later vocabularies (Weisleder & Fernald, 2013). This work offers promising tools for streamlining work with daylong recordings, facilitating research that aims to better understand how children learn from everyday speech environments.

Keywords: child-directed speech, other-directed speech, LENA, daylong recordings, automated classifier

Corresponding author(s): Janet Bang, Department of Child and Adolescent Development, San José State University, One Washington Square, San José, CA, 95192, USA. Email: janet.bang@sjsu.edu

ORCID ID(s): Janet Y. Bang: https://orcid.org/0000-0002-6014-3009
George Kachergis: https://orcid.org/0000-0003-4153-4167
Adriana Weisleder: https://orcid.org/0000-0001-6094-8424
Virginia A. Marchman: https://orcid.org/0000-0001-7183-6743

Citation: Bang, J.Y., Kachergis, G., Weisleder, A., & Marchman, V. A. (2023). An automated classifier for periods of sleep and target-child-directed speech from LENA recordings. *Language Development Research, 3*(1), 211–248. https://doi.org/10.34842/xmrq-er43

Introduction

Speech environments vary across children in numerous ways. The ability to document variation in children's naturally-occurring speech environments has been greatly assisted by technology that can capture, store, and process large amounts of audio data (e.g., an entire day). One notable example is the LENA digital language processor and software system (Gilkerson et al., 2017; Gilkerson & Richards, 2020). The recorder is worn inside a child's front shirt pocket and records the audio environment around the child, with each recording storing up to 16 hours of audio. The LENA software applies machine-learning algorithms to identify speech from children and adults that is "meaningful" or "near and clear" to the child (Cristia et al., 2021; Gilkerson & Richards, 2020). Summary reports provide estimates of the number of adult words (Adult Word Count, AWC), child vocalizations (Child Vocalization Count, CVC), and conversational turns (Conversational Turn Count, CTC), as well as the duration of time with meaningful speech, distant speech, TV/electronic media, non-speech noise (e.g., fan), and silence. A number of studies in different languages have compared these estimates to counts derived from human transcription and have reported mixed findings for the validity of LENA measures, with AWC, CTC, CVC among the most widely studied (Busch et al., 2018; Canault et al., 2016; Ferjan Ramirez et al., 2023; Gilkerson et al., 2015; Lehet et al., 2021; Soderstrom & Wittebolle, 2013; VanDam & Silbert, 2016; for a systematic review and meta-analyses of validation studies see Cristia et al., 2020).

Studies with LENA have been conducted in numerous languages and sociocultural settings. Most of these studies use LENA's estimates of AWC, CTC, and CVC to investigate how young children's language environments might support their language development, particularly by examining the amount and types of speech that are available to the child. Although the automated speech counts provided by LENA are useful, they are not sufficient to characterize many aspects of children's speech environments that are thought to be relevant for language learning. For example, segments with relatively high AWC values may indicate interactions when an adult is engaging verbally with their child (i.e., target-child-directed speech, tCDS). But these segments could also reflect periods in which multiple adults are talking to each other near the child, without any of the adults speaking directly to the child (i.e., other-directed speech, ODS). Similarly, some portions of the day may be characterized by high values for silence. These long periods of silence could reflect times when no speech is addressed to the child even though the child is awake and available to experience that speech (e.g., the caregiver is not interacting with the child or they are engaging nonverbally). Or, these periods could reflect times when the child is sleeping and no adults are present. These different scenarios have been proposed to play different roles in language learning and are of theoretical interest to many researchers. Yet, the LENA algorithms/measures do not currently distinguish between them.

Deriving estimates of the child-directed vs. other-directed nature of the speech that children hear is particularly important for our understanding of how children learn language from their speech environment (Dailey & Bergelson, 2022). A growing body

of work has proposed that target-child-directed speech, more so than other-directed speech, supports language development (Goldstein & Schwade, 2008; Shneidman & Goldin-Meadow, 2012; Weisleder & Fernald, 2013). Relatedly, when caregivers engage verbally with young children, the extent to which they use a child-directed register, i.e., speech characterized by certain acoustic, prosodic, lexical, and morphosyntactic properties, has been proposed to be particularly conducive for learning (Fernald et al., 1989; Quigley et al., 2019; Singh et al., 2009; Soderstrom, 2007; Stärk et al., 2022). These studies exemplify the rapidly growing interest in identifying and characterizing periods of target-child-directed speech within daylong recordings.

Child-directed Versus Other-directed Speech

The construct of child-directed speech is central to theories that aim to explain how children learn language from social interactions (Csibra & Gergely, 2009; Tomasello, 1995). However, communities vary widely in how much speech is directed to children and how much speech is spoken around the child but not directed to them (Casillas et al., 2019; Ochs & Schieffelin, 1984; Shneidman & Goldin-Meadow, 2012). Despite this variability, cross-cultural work finds that key language milestones (e.g., onset of first words and multi-word utterances) emerge around the same age in a variety of communities (Casillas et al., 2019; Crago et al., 1997). Such findings raise questions regarding whether any speech in children's environments, whether it is addressed to them or not, can support their language acquisition.

Indeed, lab-based experimental studies have demonstrated that children can learn new words from speech that is not explicitly directed to them. For example, Akhtar and colleagues (2001) found that 1- to 2-year-old children were able to learn novel nouns and verbs when observing two adults play a game. Other studies varied the degree of joint attention between speaker and learner, such as having speakers turn their backs to infants during a word learning episode, replicating the finding that children can learn new words even in such contexts (Gampe et al., 2012). In contrast, some research examining speech in natural environments reports that target-child-directed speech, more so than other-directed speech, is associated with children's vocabulary development (Shneidman & Goldin-Meadow, 2012; Weisleder & Fernald, 2013). For example, using LENA recordings with 29 Spanish-speaking families in the U.S., Weisleder and Fernald (2013) coded for periods of target-child-directed speech, i.e., speech directed to the target child in one-on-one interactions or with others, versus overheard speech[1], i.e., speech directed to other adults or children other than the target child. Using AWCs from LENA when the child was 19 months, they found that the number of adult words in periods with target-child-directed speech was related to children's vocabulary size at 25 months, while the number of adult words in periods

[1] Weisleder & Fernald (2013) and Shneidman & Goldin-Meadow (2012) used the term "overheard" speech. We use 'other-directed speech' as a more conservative term (Casillas et al., 2019), since it is unclear whether children do or do not hear speech when it is directed to others.

with other-directed speech was not. Similar findings were observed in Shneidman and Goldin-Meadow (2012), where the amount of target-child-directed, but not overheard, speech was associated with child vocabulary in Yucatec-Mayan-speaking families in subsistence farming communities in Mexico. Collectively, these studies reveal mixed findings about the differential roles of target-child-directed and other-directed speech in young children's language learning.

When caregivers engage with young children, they sometimes change their speech register, producing a type of speech colloquially referred to as "baby talk", "parentese", and which researchers refer to as "infant-directed speech (IDS)." Numerous acoustic, prosodic, phonological, lexical, grammatical, and pragmatic features have been noted to differentiate this child-directed register from adult-directed registers (Hilton et al., 2020; Soderstrom, 2007). Moreover, speech that is characterized by features of IDS has been suggested to be especially supportive of children's speech and language acquisition (Byers-Heinlein et al., 2021; Fernald et al., 1989; Singh et al., 2009; Snow, 1977). For example, a recent multi-continent collaboration demonstrated that speech characterized by the acoustic and phonological features of North American English IDS was preferred over speech spoken in an adult-directed register by both mono- and bilingually-exposed infants (Byers-Heinlein et al., 2021). These results were interpreted to suggest that acoustic features associated with IDS may be more effective at attracting infants' attention and thereby, can better support learning, particularly when young children are developing their early language skills. However, there is continued debate about the relative role of child- and adult-directed speech registers in children's language learning across linguistic and cultural contexts (Solomon, 2011; Cox et al., 2022).

LENA's View of the Auditory Environment

The main goal of the LENA system is to identify vocalizations from the child wearing the recorder and nearby adults, while excluding all other sounds (Gilkerson & Richards, 2020). The software uses various acoustic features to segment the audio recording and label the sounds into one of eight main categories: key child (the child wearing the recorder), adult female, adult male, electronic media (e.g., TV), other child, distant or overlapping speech, noise, and silence. The result of this process is an "Interpreted Time Segments" (ITS) file, which is in essence a diarization file (Xu et al., 2009). The ITS file is written in standard XML format and can be exported from the LENA software for each recording. The ITS file contains all the segmentation/diarization information, including the duration of each sound and its intensity (loudness).

In addition to segmenting and labeling the audio, LENA also estimates the frequency of adult words (AWC), adult-child conversational turns (CTC) and child vocalizations (CVC). To do this, LENA does not attempt to recognize actual words; instead, the algorithm estimates the number of words based on information in the speech signal, such as segment duration, syllable count, and consonant distribution (Gilkerson & Richards, 2020). These word and vocalization frequencies are estimated only from LENA's three primary speaker labels (adult female, adult male, and key child), or

what LENA calls "meaningful speech." No word/vocalization counts are estimated for other children or for distant/overlapping speech. All vocalization counts include speech-like vocalizations separated by a 300 ms break, but exclude respiratory (e.g., breathing) and digestive sounds (e.g., burping). These frequencies are exported as part of the ITS file. In addition, users can export summary-level reports from LENA, which provide word and vocalization counts (AWC, CTC, CVC) over a particular unit of time (e.g., 5 minutes, or 1 hour), as well as time-based measures of the amount of time (minutes) within that unit that contain meaningful speech (i.e., speech that is 'near and clear'), distant/overlapping speech, TV/electronic media, non-speech noise (e.g., fan), and silence. These summary reports are used by most LENA users to characterize the child's speech environment, as they provide useful information about the amount of adult speech the child hears throughout the day, the child's own vocalizations, and the number of conversational exchanges between the child and adult(s). The AWC measure is the most-widely used, as well as the most reliable/accurate of these measures. However, this measure does not distinguish whether the adult speech is directed to the child or just spoken in the child's vicinity. Additionally, LENA does not identify whether the speech is characterized by prosodic and acoustic features of child-directed register, e.g., exaggerated intonation. Thus, to date, researchers interested in these distinctions have had to rely on manual annotation.

Identifying periods of sleep, target- and other-directed speech in daylong recordings

Manual annotation of LENA recordings requires that human listeners identify periods of sleep, target-child-directed, and other-directed speech by attending to numerous cues that are available on the audio recording (Weisleder & Fernald, 2013). However, these efforts are highly labor and time intensive. Though there are emerging tools to support the rigor and efficiency of this type of manual coding (Cychosz et al., 2021; Mendoza & Fausey, 2021), efforts to automate steps in this process are also in critical need. Additionally, in some cases, ethical considerations prevent researchers from listening to the recordings (Cychosz et al., 2020).

Recent work has demonstrated progress in automating speech classifications as infant/child- vs. adult-directed registers from daylong recordings (De Palma & VanDam, 2017; Schuller et al., 2017) or laboratory stimuli (Räsänen et al., 2018; Schuster et al., 2014), mainly by focusing on the acoustic and phonetic features of speech. However, no studies to our knowledge have demonstrated the extent to which we can reliably classify whether speech was directed to the target child or not from daylong recordings, regardless of register. Thus, tools that enable classification of periods of target-child-directed and other-directed speech from features that are automatically extracted from the recordings could expand the range of cases in which such features can be examined.

Figure 1 depicts examples from three children's daylong recordings (from Weisleder & Fernald, 2013), illustrating the automated AWC estimates (adult females and adult males) per 5-min audio segment across the day. Not surprisingly, the AWC values in

each segment for a given child vary considerably across the day, and the mean AWC values that are averaged across the day also vary across the three children. To determine which AWC values reflect tCDS rather than other-directed speech, human listeners judged each 5-minute segment first as whether the child was sleeping and, if not, whether the adult speech during the segment was more than 50% tCDS or ODS. Notably, removing ODS segments changed the estimates of overall speech to the child across the day substantially for some children, but less so for others.

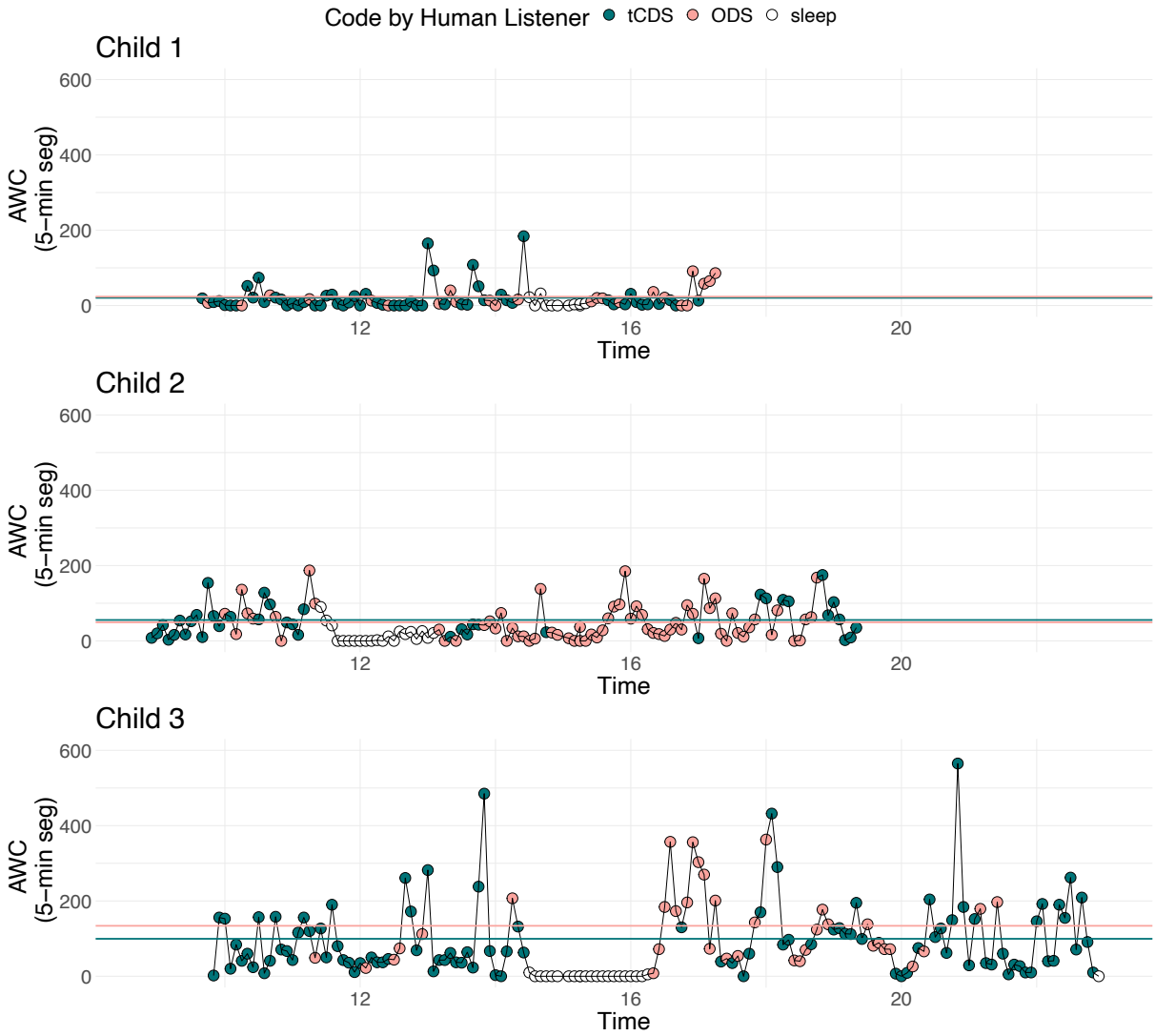


Figure 1. ***Example profiles of three children's AWC counts per 5-minute segment across their daylong recordings. Note: Green dots represent segments judged by human listeners to be more than 50% target-child-directed speech; Light pink dots represent segments judged to be more than 50% other-directed speech; White dots represent segments when the child was sleeping as judged by human listeners. Horizontal lines depict the average tCDS (green) or ODS (pink) counts computed over the entire recording.***

As this figure shows, periods of tCDS or ODS were not easily differentiated by AWC (i.e., segments that were identified as tCDS and ODS had a range of both high and low AWC values). Thus, to differentiate periods of tCDS and ODS, it may be more productive to examine multiple measures in combination. For example, a given 5-minute segment may be more likely to be tCDS when that segment's AWC value is interpreted in conjunction with relatively high values of CTC or CVC. Similarly, one could predict that periods of children sleeping would be characterized by both low values of AWC and low values of CVC or CTC. By combining across measures, we can gain insights into which features conspire to reflect periods of sleep or of tCDS versus ODS and how best to identify them automatically.

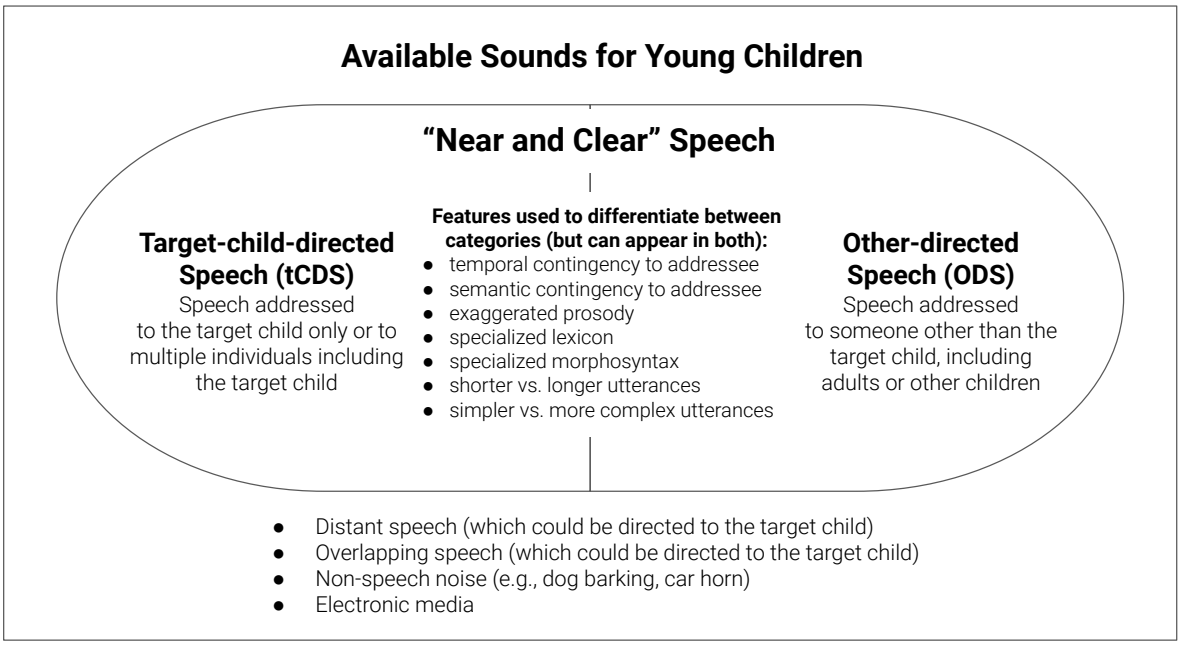
Current Study

This study examined ways to facilitate the identification of periods of target-child-directed vs. other-directed speech in daylong LENA recordings, as well as periods when children are awake versus sleeping, using only the automated measures provided in the standard LENA summary reports. By focusing on the automatically-generated LENA measures, we seek to develop classification tools that require minimal additional processing of the data and that can be easily integrated into a workflow.

Figure 2 provides our conceptualization of target-child-directed vs. other-directed speech. Looking only at the speech that is "near and clear," i.e., potentially audible by the target child, we defined tCDS as all speech that is directed to the target child, either individually or part of a group. In contrast, ODS is defined as all "near and clear" speech that is addressed to others. Note that other features cross-cut these categories. For example, while tCDS is more likely to be characterized by short utterances and child–directed prosody, there are times when speech that is clearly directed to a child does not fit that characterization. Analogously, while ODS may be more likely to be spoken in adult-directed prosody, there are also times and contexts when ODS might share many of the features, e.g. exaggerated prosody, characteristic of child-directed speech. Our goal was to develop a tool that could effectively identify periods of all speech directed to the target child, some of which may use a CDS register and some which may not. By focusing more generally on the function, rather than the features of speech, we align our research questions with the more general theoretical proposal that children learn language through interactions with others, and that they may learn best from language input that is contingent on or relevant to their vocalizations, actions, and/or attentional focus (Goldstein & Schwade, 2008; McGillion et al., 2013; Tamis-LeMonda et al., 2014; Tomasello, 1995; Yurovsky, 2018).

Our approach is as follows. We first conducted preliminary analyses to explore how the core frequency count measures, i.e., AWC, CTC, and CVC, worked in combination to predict periods of target-child-directed vs. other-directed speech. Using data from recordings of 29 Spanish-speaking families in the U.S. (from Weisleder & Fernald, 2013), we conducted logistic regressions to assess the degree to which variation in these measures was associated with whether a particular 5-minute segment was classified as tCDS or ODS by human coders. Next, we compiled data across several studies

of English- and Spanish-speaking families in the U.S. (n = 153), applying more complex machine-learning classifiers that combined the frequency (AWC, CTC, CVC) and time-based measures (meaningful speech, distant speech, TV, noise, and silence) to



identify periods of sleep, tCDS, and ODS that had been previously identified by human coders. We first used cluster analyses to examine how these multiple LENA features hung together and then developed two classifiers, one for distinguishing periods when the target child was asleep versus awake and another for distinguishing periods of primarily tCDS versus ODS.

Figure 2. ***Conceptualization of tCDS and ODS in our study. Note: "Near and clear" describes the audible speech from the perspective of the child wearing the recorder (Cristia et al., 2021; Gilkerson & Richards, 2020), which we define as the target child.***

Performance of both the sleep and tCDS/ODS classifier were evaluated based on the concordance with the human coders, defined in terms of both the sensitivity and specificity of the model predictions in comparison to human coders (ground truth). These estimates provide a standard measure of classification ability reflecting the degree to which the classifiers can distinguish both negative and positive values of each category. For the tCDS/ODS classifier, we also evaluated its performance in terms of its ability to replicate previously published links between variation in adult word counts and children's later vocabulary outcomes. In particular, Weisleder & Fernald (2013) reported stronger correlations between parent-reported vocabulary size and AWC values derived from 5-minute segments categorized as primarily tCDS, compared to those based on 5-minute segments identified as being primarily ODS. If a similar pattern of correlations is found with classifier-based values, this would provide some assurance that the classifier is capturing dimensions of children's language

input that are analogous to those identified by human coders.

Methods

Participants

Participants were families and their 17- to 28-month-old children from 79 English- and 74 Spanish-speaking households in the U.S. In total, families contributed over 1,000 recorded hours of LENA recordings (12,936 5-min segments). Descriptives are shown in Table 1. Data analyzed were collected between 2008-2015. Recruitment information is reported elsewhere (Fernald et al., 2013; Marchman et al., 2020; Weisleder & Fernald, 2013).

Table 1. ***Descriptive statistics of participants and recordings in the five different samples.***

Sample	n	Lang.	Age range (mo)	Mat. Ed range (y)	Total recording length in hours Mean (SD)	Seg. dur (min)	Num seg. incl.
1	27	En	18 - 19	12 - 18	10.62 (2.29)	5	3491
2	29*	En	17 - 19	10 - 18	9.32 (2.52)	5	3275
3	45	En	23 - 26	10 - 18	11.05 (3.22)	5**	1891
4	29	Sp	18 - 20	4 - 16	10.67 (3.13)	5	2758
5	45	Sp	25 - 28	6 - 18	13.44 (3.68)	5**	1521

Note: *n = 22 from Sample 2 are also included in Sample 3 at a second time point, thus the total sample results in 153 unique families; En = English, Sp = Spanish. **10-minute segments rated by human coders were split into 5-minute segments for the purpose of our analyses.

Data collection and Coding

Across all studies, research staff obtained informed consent from caregivers and provided instructions of how to use LENA. Caregivers were asked to record on a "typical day." To respect families' privacy, caregivers were told that they could pause the recording at their convenience. Recording instructions varied slightly across samples, but in all cases, families were given a single LENA recorder to use on a single day or across multiple days. All families were encouraged to record during all parts of the day. All recordings were cleaned following a standard lab protocol to exclude portions

of the recording when the LENA was not being used as recommended (e.g., the child was not wearing the vest, or the caregiver asked us not to listen to a period of the day.) Details about cleaned versus uncleaned recordings can be seen in Bang et al. (2022) and Weisleder & Fernald (2013).

Next, native speakers of each language coded segments of the audio-recording. For all samples, coders classified each segment as tCDS or ODS based on the most prevalent type of speech in that segment. For samples 1, 2, and 4 (see Table 1), human listeners listened to the entire recording and coded each 5 min segment as consisting of: sleep, primarily tCDS, primarily ODS, or a 50/50 split between tCDS or ODS. Segments of sleep were confirmed by environmental sounds (e.g., deep breathing). Segments identified as tCDS were those in which the majority (> 50%) of the surrounding adult speech (i.e., represented by the AWC value) was directed to the target child wearing the recorder, either addressed exclusively to the target child or inclusive of the target child (e.g., a speaker addressed a group that included the target child). Coders based their judgments on numerous features including the content of the speech, as well as exaggerated prosody, slower speech tempo, affect, perceived distance of the speaker relative to the child, environmental sounds, who responded to the speaker, and the activity of the interaction. Segments identified as ODS were those in which the majority of the speech was not directed to the target child nor inclusive of the target child. Split segments were those judged to have equal amounts of tCDS and ODS. For all preliminary analyses and the classifiers, we treated all 'split' segments as ODS, so that all segments coded as tCDS reflected segments with more than 50% target-child-directed speech.

For samples 3 and 5, a slightly revised protocol was followed. Here, coders first listened to potential periods of sleep based on information in a log book, targeting segments with consecutive low AWC values (AWC values = 0 for a minimum of 2 consecutive segments). If the child was confirmed to be sleeping, coders continued listening to segments prior to and after these segments to determine the beginning and end of periods of sleep. Next, families' highest AWC values were sorted in descending order based on 10-min segments, and coders rated each segment as primarily tCDS or ODS if approximately 70% of speech was either tCDS or ODS until six segments of primarily tCDS were identified per family (Bang et al., 2022). These 10-min segments were split into 5-min segments for the purpose of the current analyses, attributing the assigned code to each of the 5-min segments.

Reliability

To assess reliability of human coding, we determined the degree to which judgments of tCDS or ODS were consistent between two human raters. For each sample, we randomly selected 5 families (approximately 10 - 20% per sample, depending on the sample size) to be double-coded. For samples 1, 2, and 4, each family's recording was split into thirds and we randomly sampled five continuous 5-min segments from each block. Continuous segments were selected for double-coding in order to create coding

conditions that were analogous to those of the original coders who listened to the entire recording. For samples 3 and 5, we randomly selected ~eight 10-min segments for two families per sample, splitting the 10-min segments into 5-min segments, for a total of ~16 5-min segments coded by second raters. For purposes of reliability calculations, we excluded segments identified as splits during initial coding (n = 37, 9.7% for samples 1, 2, and 4) and other segments that were previously removed from analysis (n = 15, 3.9%). Judgments were compared between two raters (with different combinations of first and second raters), using K = 2 codes (tCDS or ODS), and raters coded independently (i.e., second raters had no knowledge of codes by first raters). Our coding protocol can be seen here: https://osf.io/qcj6r/.

For all samples, first raters were treated as the gold standard. Human raters judged each 5-min segment as having (a) no caregiver speech, (b) <less than 50% tCDS, (c) 50 - 70% tCDS, or (d) >70% tCDS. Segments rated as (a) or (b) were considered ODS; segments rated as (c) or (d) were considered tCDS. We evaluated our interrater reliability using Cohen's kappa and estimated rater accuracy (Bakeman, 2022). The value of "estimated rater accuracy" is determined from a simulation using the KappaAcc program (Bakeman, 2022), and reflects how accurate simulated observers would need to be to obtain the same observed kappa given the specifics of the data. Estimated rater accuracy provides a metric to judge "accurate enough" standards given the conditions of a dataset (e.g., number of raters and codes, frequency of different codes), rather than categorical cutoff points for Cohen's kappa. Table 2 reports that our Cohen's kappa for the total sample was .54 (80% agreement, uncorrected for chance). To produce a kappa of this value, the estimated rater accuracy suggests that simulated observers under similar circumstances (2 codes, 2 raters) would need to be 87% accurate (range across five samples = 77 - 90%).

Table 2. ***Reliability between first raters and second raters per sample and in total***

Sample	Language	n	Percent agreement (uncorrected)	Estimated rater accuracy	Cohen's kappa
1	En	5	85%	87%	.38
2	En	5	80%	89%	.61
3	En	5	79%	88%	.58
4	Sp	5	83%	90%	.65
5	Sp	5	73%	77%	.24
Total	En and Sp	25	80%	87%	.54

Results

Preliminary Analyses

Figure 3 illustrates the distributions of raw AWC, CTC, and CVC values per 5-min segment for tCDS or ODS segments using only data from Sample 4 (Weisleder & Fernald, 2013). To examine the degree to which the frequency measures of AWC, CTC, and CVC predicted the human-coded classifications of tCDS or ODS, we conducted hierarchical mixed effects logistic regression models. Models included a random intercept per participant and importantly, all frequency measures were converted to rates per minute and mean-centered within each family to allow interpretation of values as relative to each family's mean rates.

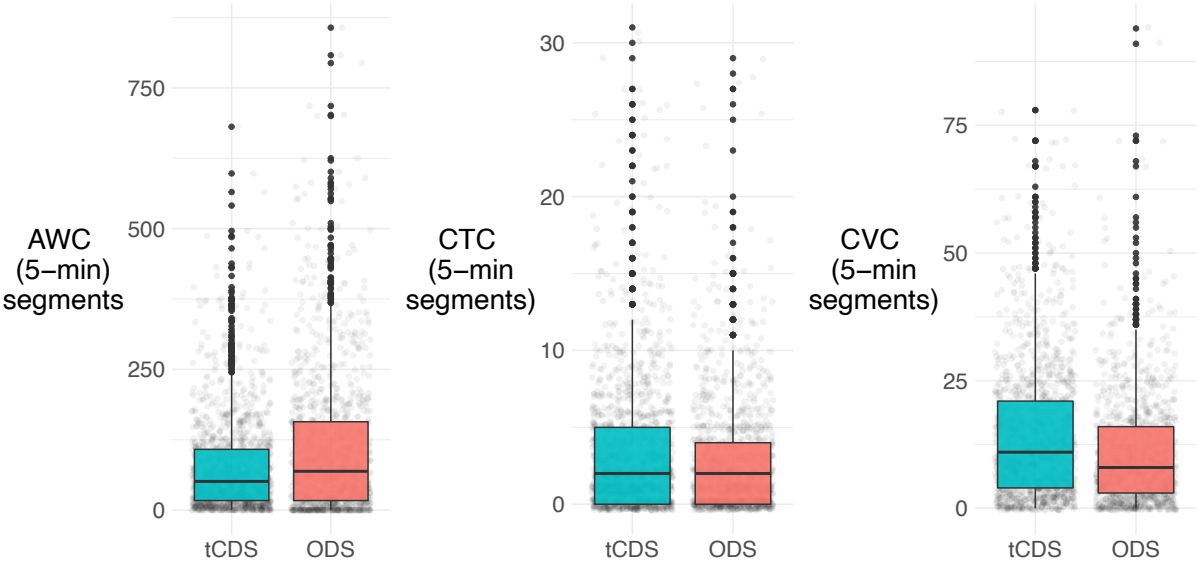


Figure 3. ***Boxplots of raw AWC, CTC, and CVC values by 5-min segments human-coded as tCDS or ODS using data from Sample 4. Note: Split segments were treated as ODS; segments identified as sleep were excluded.***

We found that each frequency measure, AWC/min, CTC/min, and CVC/min, independently contributed to the probability of a segment being classified as tCDS versus ODS. As seen in Figure 4, lower AWC rates (B = -.59, 95% CI = [-.72, -.46]) were associated with a higher probability of tCDS, indicating that segments that have higher than average AWC for a given family have a lower probability of being coded as tCDS by human coders. In contrast, higher rates of CTC (B = .36, 95% CI = [.21, .52]) and CVC (B = .39, 95% CI = [.26, .52]) were associated with a higher probability of tCDS, such that segments that were higher than average in CTC or CVC for a given family were more likely to be coded as tCDS. These findings indicate that each of the LENA frequency measures predicted the probability of tCDS, but did so in different directions. Moreover, because these measures are interrelated, it is likely that relations were more complex than these techniques could capture. Thus, we next recruited machine

learning techniques to explore the extent to which multiple LENA features, including both frequency- (AWC, CTC, and CVC) and time-based (e.g., minutes in meaningful speech, noise), could be used jointly to classify periods of sleep, tCDS or ODS.

Cluster Analyses

We next examined whether segments could be meaningfully clustered, which might suggest that a classifier based on thresholding multiple feature values (e.g., a decision tree) might work better than techniques that looked at predictors individually. We include speech frequency measures (AWC, CTC, and CVC) and time-based measures provided by LENA summary outputs (minutes in meaningful speech, distant speech, TV, noise, and silence), and examined how these measures clustered to predict human coding of the 5-min segment as periods of sleep, >50% tCDS, or > 50% ODS. Using an unsupervised clustering algorithm (k-means), we clustered all 12,936 segments according to their raw LENA values, considering k = {2,..,15} clusters. Table 3 shows the selected k = 7 clusters along with the proportion of each type of segment in the cluster and the mean values of LENA features for segments in that cluster. As shown in bold, Clusters 4 and 5 capture mostly sleep (64% and 53%) with low AWC, CTC, and CVC, but both clusters also include a moderate number of tCDS segments (22% and 30%). Note that Cluster 5 is also associated with high levels of noise (italicized), whereas Cluster 4 is associated with high levels of silence. The next two clusters in bold, Clusters 6 and 1, are both predominantly tCDS (73% and 60%) and cover 36.4% of the dataset, however, one has somewhat higher mean AWC, CTC, and CVC values than the other. Note also that these two clusters also contain many ODS segments. Next, we can note that Clusters 7 and 2 are comprised mostly of ODS segments. While both clusters are associated with low values of CTC and CVC, Cluster 7 is associated with high values of AWC, while Cluster 2 is not. Finally, Cluster 3, which looks much like the sleep clusters (4 and 5) in terms of low AWC, CTC, and CVC, is also associated with a higher level of TV than other clusters.

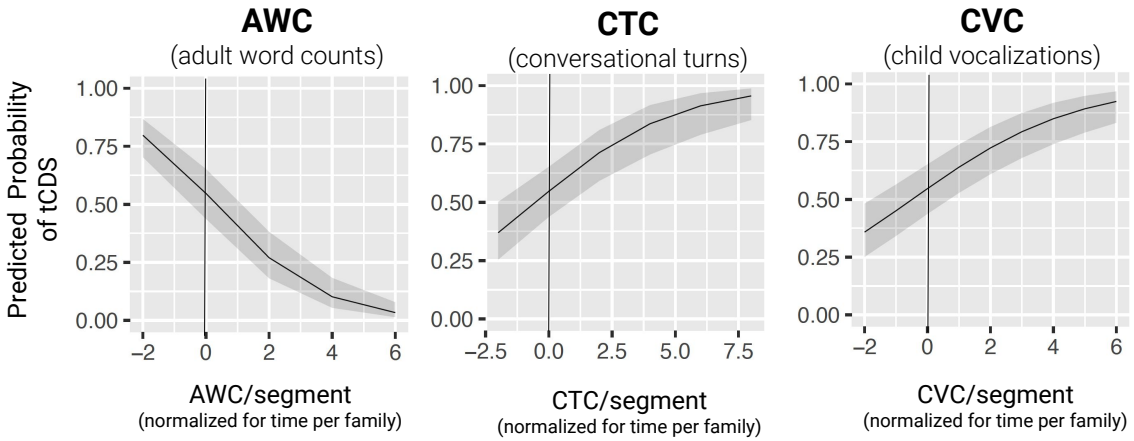


Figure 4. ***Predicted probabilities and confidence intervals (shaded region) of tCDS from AWC, CTC, and CVC, when holding each other measure at families' mean value (vertical line at 0).***

Overall, these cluster analyses showed that: 1) multiple LENA features captured meaningful variation between the clusters, as some features clustered together to correspond primarily to sleep, tCDS, or ODS, and yet 2) the clusters have significant overlap in tCDS and ODS, and to a lesser extent, sleep.

Table 3. ***Means of LENA features by cluster, annotated with proportion of sleep, tCDS, and ODS segments.***

		Category			LENA Features							
cluster	**N**	**sleep**	**tCDS**	**ODS**	**AWC**	**CTC**	**CVC**	**noise**	**silence**	**distant**	**TV**	**meaningful**
4	2,041	**0.64**	0.22	0.14	3.0	0.1	0.5	0.01	*0.85*	0.08	0.02	0.03
5	142	**0.53**	0.30	0.17	3.4	0.1	0.9	*0.63*	0.12	0.16	0.04	0.04
6	1,256	0.00	**0.73**	0.27	54.6	*3.8*	*9.5*	0.02	0.27	0.25	0.01	*0.45*
1	3,450	0.01	**0.60**	0.39	22.0	1.1	4.8	0.03	0.37	0.33	0.03	0.25
7	1,485	0.01	0.33	**0.66**	*76.1*	1.4	2.6	0.01	0.21	0.33	0.03	*0.42*
2	3,475	0.04	0.45	**0.51**	13.6	0.4	1.8	0.03	0.17	*0.66*	0.02	0.12
3	1,087	0.27	0.28	**0.45**	7.3	0.2	0.7	0.02	0.15	0.07	*0.69*	0.06

Note: Bolded numbers correspond to clusters that included the highest proportion of segments classified most frequently as sleep, tCDS, and ODS, respectively. Italicized values indicate maximum cluster means of each LENA feature. AWC, CTC, CVC are automated counts per 5-minute segment, normalized to be rates (counts per minute). Values for noise, silence, distant, TV, and meaningful are proportions of each per 5-minute segment.

Classifying Sleep Segments

We attempted to build a classifier to automatically distinguish sleep from awake segments using only automatically-generated LENA features. All counts and durations of time were normalized to per-minute values (i.e., divided by segment duration). Although we experimented with simpler classification algorithms (e.g., decision trees and random forests; Bang et al., 2022), ultimately the best performance was achieved with XGBoost (eXtreme Gradient-Boosted trees; (Chen & Guestrin, 2016), a state-of-the-art algorithm that trains a cascade of decision trees successively on subsets of the data, upweighting the segments that were misclassified by earlier decision trees.[2] It

[2] XGBoost takes an MxN matrix of M training samples (5-minute segments, in our case) of N numeric features (scaled LENA metrics, here), and iteratively constructs a set of decision trees that aim to predict the given binary classes (e.g., sleep / not-sleep; or CDS / non-CDS), where each new tree focuses more on the data points that were misclassified by prior trees.

should be noted that XGBoost does not work well in some domains (e.g., it does not appear to work well for object recognition in images, Ohn-Bar & Trivedi, 2016), and tree-based methods in general do not extrapolate well beyond the range of feature values in the training set. Thus, it is important to thoroughly test via cross-validation, and to have a large and diverse training set to improve generalizability. An XGBoost classifier was trained using the xgboost R package (v1.7.5.1; Chen & He, 2023) to distinguish segments when the target child was asleep from those when they were awake, mirroring the first step that researchers could take when manually cleaning a LENA dataset. We trained the model using 5-fold cross-validation on 90% of the dataset (11,642 of 12,936 segments) and then tested the model on the remaining 10% held-out data (1,294 segments).

Results for the held-out data of the cross-validated classifier are shown in Figure 5. We illustrate the Receiver Operating Characteristic (ROC) curve, which depicts the performance of the classifier on sensitivity vs. specificity given all discrimination threshold values. On the left, the ROC curve reflects an overall ratio of sensitivity (y-axis) to specificity (x-axis) that was quite high, an Area Under the Curve (AUC) > .95, on the held-out test segments, with an accuracy of 0.945.[3] One limitation of XGBoost is that it does not enable simple visualizations, e.g., a decision tree, of how classifications are made. However, the feature importance measure can be used to assess which features were most informative in the ensemble of boosted trees. Shown in the right-hand panel of Figure 5, the amount of meaningful speech was the most important feature for classifying sleep segments, followed by the amount of silence, the number of child vocalizations, and distant speech.

A final sleep classifier was trained using all of the data (12,936 segments; 1,879 sleep segments, 11,057 awake), resulting in a classifier with superior performance to the cross-validated classifier (AUC = 0.985; see Appendix A for additional details). This sleep classifier has been made accessible for other researchers in a web app.[4]

Classifying tCDS vs. ODS Segments

We turn now to the more challenging task of building a classifier to automatically distinguish tCDS from ODS segments. We trained an XGBoost classifier on LENA features to distinguish tCDS segments from all other segments (ODS and split segments). First, we removed the 1,879 human-coded segments during which children were asleep (assuming they would be removed manually or by the sleep classifier). We then reclassified the 1,012 "split" segments which human coders judged to be 50% ODS and 50% tCDS as ODS, resulting in a total of 5,239 ODS segments and 5,818 tCDS segments

[3] To test whether the classifier was overfitting to characteristics of particular segments, we trained a 5-fold cross-validated version on 80% of the children, leaving out data from 20% of the children (n=30) in each fold. This classifier achieved very similar performance (AUC = 0.95; average test accuracy = 0.95), suggesting that the classifier will generalize to new children from similar samples. ROC curves for this analysis are shown in Appendix F.

[4] https://kachergis.shinyapps.io/classify_cds_ods/

(58% tCDS) when children were awake. The purpose of the classifier is thus to distinguish periods with >50% tCDS from segments that were at least 50% ODS, after removing periods of sleep. A random 90% of the awake data (9,951 out of 11,057 segments) was used to train the classifier, and the remaining 10% served as the held-out test set (1,106 segments) for evaluation.

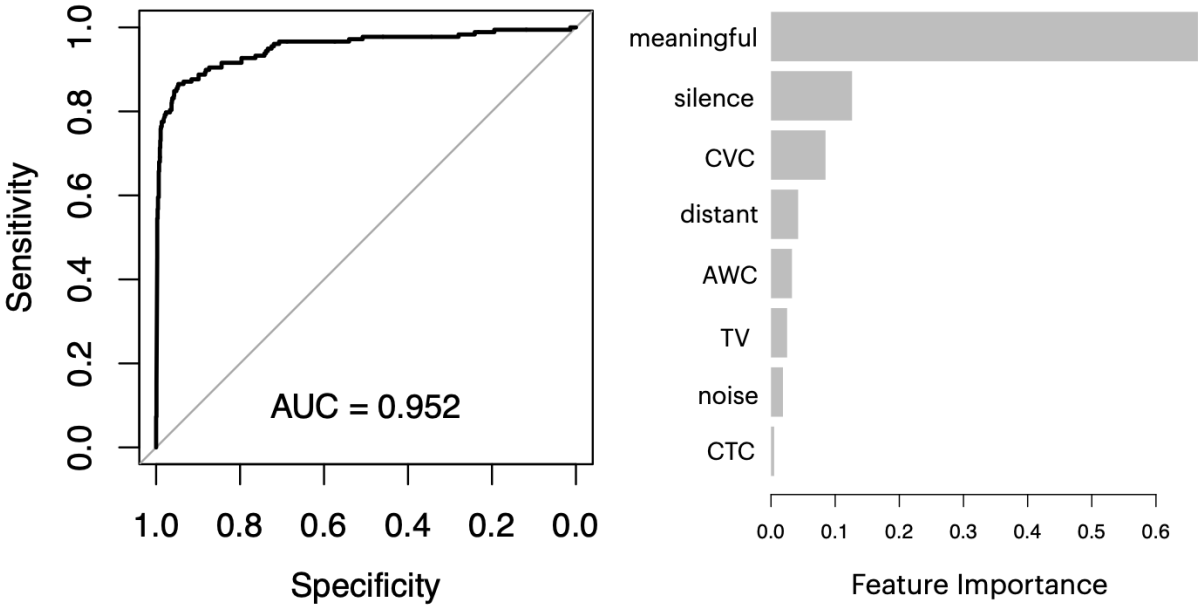


Figure 5. ***(left) The ROC curve for the sleep classifier for the 10% held-out test set. (right) Relative importance of the LENA features in the XGBoost sleep classifier trained on 90% of the data.***

The results are presented in Figure 6. As shown in the left-hand panel, when trained on 90% of the segments, the XGBoost classifier achieved moderate overall classification performance (AUC = 0.719), with an overall accuracy of 0.674 on the held-out data.[5] As shown in Figure 6 (right), the four most important features were the duration of silence, CTC, AWC, and meaningful speech.

A final XGboost classifier was trained with all 11,057 segments in order to offer the best chance for generalization to new data with similar samples, though there is no guarantee of similar performance for families and settings dissimilar to the present dataset. This final classifier's performance is shown in the Appendix Figure B1 and in

[5] To ensure that the classifier was not overfitting to these segments, we also trained a cross-validated version on 80% of the children, leaving out data from 20% of the children (n = 30) in each fold. This classifier achieved approximately the same performance (AUC = 0.73; average test accuracy = 0.66), suggesting that the classifier will generalize similarly well to data from additional children (see Appendix F for ROC curves). We also investigated including demographic and time of day features in the classifier, but found that inclusion of these features resulted in overfitting (i.e., poorer performance on held-out data).

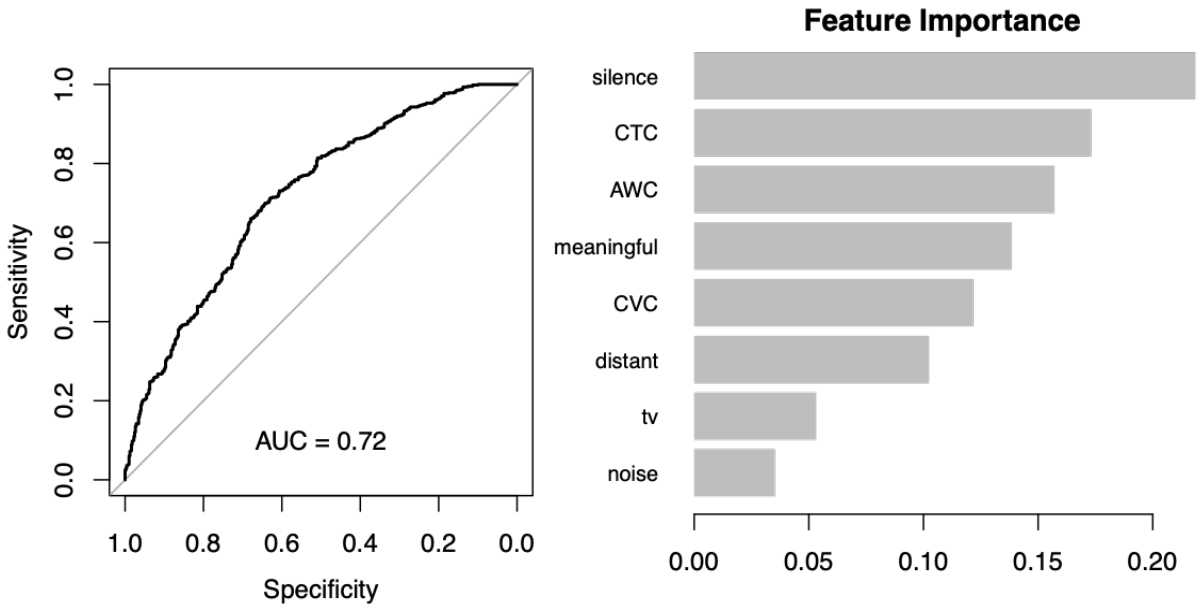


Figure 6. ***(left) ROC curve of the tCDS/ODS classifier for the 10% held-out test set and (right) relative importance of the LENA features in XGboost classifier trained on 90% of the data.***

the confusion matrices in Table 4. The AUC for this final classifier is much improved (AUC = 0.83), but performance on new data may be expected to be in-between the 90%-trained classifier and this higher value. This tCDS/ODS classifier is available for other researchers to use via a web app.[6]

Reliability Between the tCDS/ODS Classifier and a Human Rater

How does our classifier compare against human raters? Table 4a shows the confusion matrix for agreements (diagonal) and disagreements (off-diagonal) between the human raters (row) and the classifier's (columns) final binary predictions. The classifier correctly identified 80% of segments that humans rated as tCDS, as well as 70% of segments that humans rated as ODS. For comparison, Table 4b shows the confusion matrix for agreements (diagonal) and disagreements (off-diagonal) between two human raters. On average, human raters had 87% agreement for tCDS and 65% agreement for ODS. Thus, while tCDS agreements were slightly higher between two human raters and ODS agreements were slightly stronger between a classifier and a human rater, both confusion matrices indicate that ratings were similar whether comparing the classifier against a human rater or between two human raters. Sample-specific results can be seen in Appendix C.

[6] https://kachergis.shinyapps.io/classify_cds_ods/

Table 4. ***Confusion matrices.***

a) Human rater 1 (Gold Standard) vs. tCDS/ODS classifier

		Classifier		
		tCDS	ODS	Total
Human rater (Gold Standard)	tCDS	4641 (80% agreement)	1177	5818
	ODS	1554	3685 (70% agreement)	5239
	Total	6195	4862	11,057

Note: a) The diagonal (gray shading) indicates agreement between a human rater and the final XGboost classifier. b) The diagonal indicates agreement between two human raters. There were multiple individuals who served as first and second raters. For both tables, the percent agreement is calculated by dividing the number of agreements by the gold standard's total codes of the respective category.

b) Human rater 1 (Gold Standard) vs. Human rater 2

		Human rater 2		
		tCDS	ODS	Total
Human rater 1 (Gold Standard)	tCDS	190 (87% agreement)	28	218
	ODS	39	74 (65% agreement)	113
	Total	229	102	331

Links Between tCDS and Child Language Outcomes

One critical question is whether the tCDS/ODS classifier works sufficiently well to replicate results from studies with human-coded data. To test this, we used the Weisleder & Fernald (2013) dataset of 29 Spanish-speaking children whose caregivers completed the MacArthur-Bates Mexican Spanish CDI (Jackson-Maldonado et al., 2003) to assess vocabulary size when the children were 24 months. As illustrated in the left-hand panel of Figure 7, in this human-coded dataset, children who heard more tCDS at 19 months had significantly larger vocabularies at 24 months (r = .52, 95% CI = [.19, .75], p = .004). However, there was no significant association between the amount of ODS at 19 months and vocabulary size at 24 months (r = .25, p = .199).

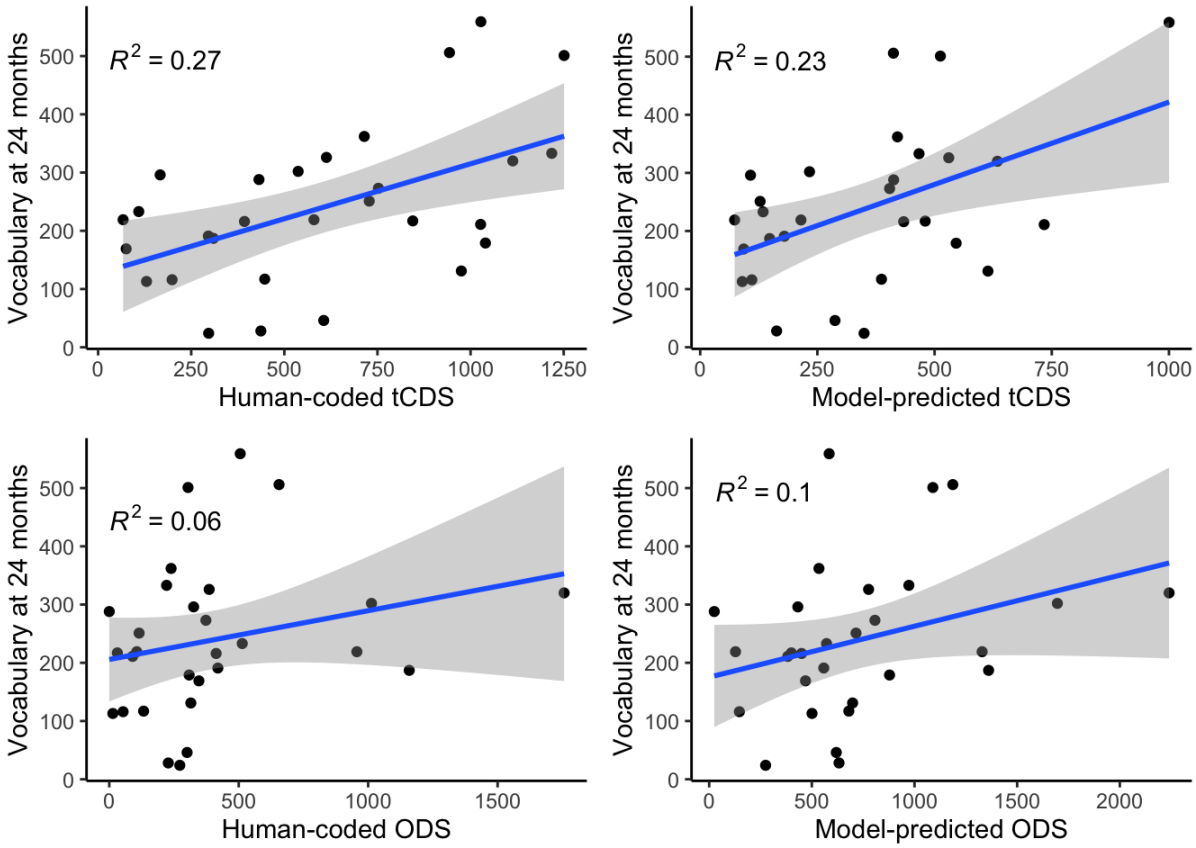


Figure 7. ***Scatterplots between human-coded or model-predicted tCDS or ODS at 19 months and children's later vocabulary sizes at 24 months. Note: Associations between vocabulary size and tCDS tokens are significantly positive, and of similar magnitude, whether human-coded (top, left) or model-predicted (top, right). Associations between vocabulary size and ODS tokens are not significant, but are of similar size, both for human-coded (bottom, left) and model-predicted (bottom, right) segments.***

We investigated these same correlations using the classifier's predictions of which segments were classified as tCDS vs. ODS. As shown in the right-hand panel of Figure

7, as with the original manual annotations, children who heard more tCDS at 19 months had significantly larger vocabularies at 24 months (r = .48, 95% CI = [.14, .72], p = .008), while the relation between the amount of ODS and vocabulary size was smaller, and did not achieve statistical significance by standard conventions (r = .32, 95% CI = [-.06, .61], p = .094). Notably, the pattern of the strength of the correlations are similar between the human-coded and model-predicted classifications, suggesting that the classifier is an effective tool for this purpose. To test whether this result was due to the inclusion of the Weisleder & Fernald dataset in the classifier's training set, we trained a classifier excluding this dataset, and found similar a pattern of results (tCDS vs. vocabulary r = .44, 95% CI = [.08,.69], p = .018; ODS vs. vocabulary r = .33, 95% CI = [−.04,.62], p = .082).

Leveraging Classifier Confidence

Although the classifier's binary performance is significantly above chance, there is substantial room for improvement, and thus we explored a more fine-grained measure of performance to determine whether some segments should be further examined by human coders. The source of XGboost's binary distinction between tCDS and ODS is actually a probability of tCDS in the range of [0,1], thresholded at 0.5 (i.e., if Pr(tCDS) > 0.5, a segment is classified as tCDS; otherwise it is classified as ODS). Figure 8 shows a histogram of classifier ratings (Pr(tCDS)) for all 11,057 awake segments in our full sample, color-coded by the classifications given by human listeners, with a dashed line indicating the threshold used for binary classification. Notably, there is significant overlap of the two distributions: there are many tCDS segments that (to the classifier) resemble and are thus confusable with ODS segments, and vice-versa. Of the 3,136 segments that were classified as tCDS with what could be considered to be low-confidence (0.4 < Pr(tCDS) < 0.6), 49% of them were judged by human coders to be tCDS. In contrast, a larger fraction of the segments classified with high confidence by the classifier agree with the human coder classification: for example, 89% of the 2,884 segments rated as Pr(tCDS) > 0.7 were judged to be tCDS by human coders, and 88% of the 2,196 segments rated as Pr(tCDS) < 0.3 were judged to be ODS by human coders. Thus, the probability of a segment being classified as tCDS could be used by researchers to make decisions about future coding or analysis, a point we return to in the discussion.

General Discussion

Our study suggests that a combination of automatically-generated measures of children's speech environments from LENA can be used to identify periods of sleep, tCDS, and ODS in daylong audio recordings, thus facilitating investigation of potentially meaningful sources of variation in young children's speech environments. We discuss our five main insights in turn.

First, we found differences in how the commonly-used, core frequency measures from LENA (AWC, CTC, and CVC) predicted the probability of a 5-minute segment being classified as containing primarily target-child-directed versus other-directed

speech. Our preliminary analyses indicated that segments with higher AWC relative to a family's mean were more likely to be judged by humans as having primarily other-directed speech. Frequency measures of CTC and CVC resulted in the opposite prediction, where segments with higher values relative to a family's mean were more likely to be judged as having predominantly target-child-directed speech. These findings suggest that periods of speech directed to a target child are defined by relatively lower rates of adult words and relatively higher rates of conversational turns and child vocalizations. This is consistent with the finding that adults often use a slower speech-rate when talking with children and that target-child-directed speech is more likely to elicit vocalizations from the child than other-directed speech. This finding also suggests that one reason some studies have found LENA's CTC measure to be a better predictor of child language outcomes than AWC (Gilkerson et al., 2018; Romeo et al., 2018) may be that high CTC is a better indicator of periods with target-child-directed speech than is AWC.

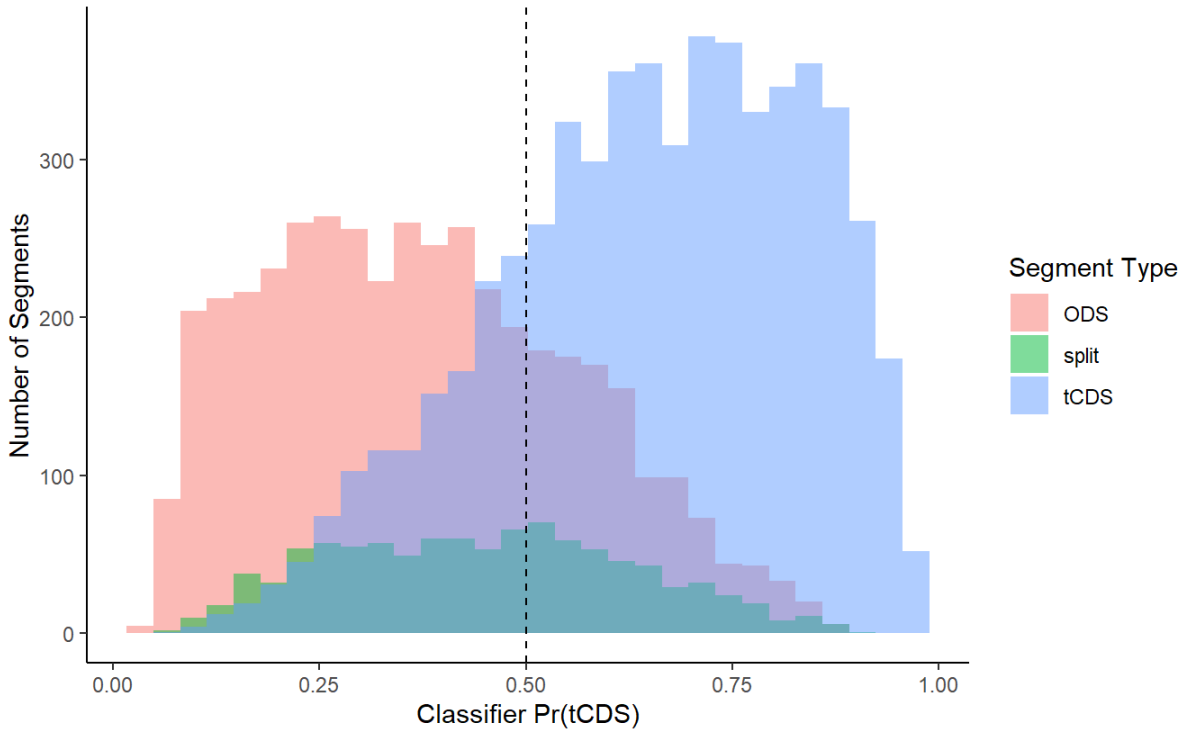


Figure 8. ***Histogram of classifier Pr(tCDS) for each segment, colored by human-coded segment type. Note: Dashed line indicates the threshold for binary classification: segments to the right were human-coded as tCDS (blue), while those to the left were human-coded as ODS (pink). Note that 'split' segments (green), which human coders found to be a mixture of both tCDS and ODS, were also given less decisive ratings of Pr(tCDS) by the classifier. The purple area indicates the overlap between tCDS and ODS regions.***

Second, a much more complex picture arose when including both LENA frequency

and duration measures in cluster analyses. While some distinct features characterized different audio environments, there was also a high degree of overlap across clusters. For example, as expected, clusters with more sleep segments were characterized by the lowest rates of AWC, CTC, and CVC. However, one sleep cluster was characterized by more silence, while the other was characterized by more noise. This aligns with anecdotal reports by human coders that periods of sleep sometimes involved what appeared to be fans or sound machines, sounds which were likely categorized as "noise" by LENA. Baby snores, which also sometimes occurred during periods of sleep, could also have been categorized as "noise" by LENA. In contrast, those clusters that were likely to be tCDS were characterized by the highest averages of CTC and CVC, but were more mixed with regards to AWC. Of clusters likely to be ODS, one cluster consisted of the highest average AWC, while the others had lower CTC and CVC rates, or longer durations of distant speech and TV. Thus, we observed multiple ways in which features were combined in clusters of predominantly sleep, tCDS, and ODS. Moreover, in no cases were sleep, tCDS, or ODS associated with only one cluster or configuration of features. Future work might fruitfully examine in more detail the potential differences between segments in different cluster types. For example, are segments in some clusters associated with different types of language interactions and/or activities than other segments?

Third, we found a high degree of success in training a classifier to identify periods of sleep in our dataset. Consistent with the multifaceted nature of clusters defined by more sleep, the classification was not simply due to periods of silence. The classifier mostly relied on the duration of 'meaningful' speech, followed by the duration of silence, and the number of vocalizations by the target child. This suggests that, at least among English- and Spanish-speaking families in the U.S., periods in which the target child is asleep vs. awake could be reliably identified from characteristics of the audio environment and shows advantages of considering multiple features of those environments.

Fourth, we found moderate success in training an XGBoost classifier to distinguish periods of tCDS versus ODS in our dataset. We found moderate sensitivity and specificity on the full dataset and a slightly smaller AUC on the held-out test segments. The feature importance list illustrated the average gain in our prediction of tCDS versus ODS, highlighting many features (meaningful speech, AWC, CTC, and silence) that also emerged in our cluster analysis. Reliability between two trained human raters suggests that even when individuals undergo training and interpret all available information in the auditory environment, there is variability across samples and there may be a ceiling of 'good enough' reliability. The moderate success of the classifier in terms of sensitivity and specificity, as well as performance seen in the confusion matrices, were similar to that of two human raters. This suggests that the level of accuracy achieved by the classifier may be a reasonable goal given the complexities of the speech environment. The superior performance of the classifier relative to analyses that were limited to individual predictors (i.e., the logistic regressions presented in our first analysis) suggests that human classifications of target-child-directed and

other-directed speech rely on nuanced distinctions that take into account combinations of features in the audio environment (e.g., low silence with high CTC and moderate AWC), as well as features of the environment not captured by these measures (e.g., semantic content).

Finally, we demonstrated that we could use model-derived predictions of tCDS and ODS to replicate associations between caregiver speech at 19 months and children's vocabularies at 24 months that were observed in previously published work in Spanish-speaking families in the U.S. (Weisleder & Fernald, 2013). We examined these correlations to test the performance of the classifier and not as an extension of the original study. The model-predicted classifications revealed, as observed with human-coding, that variability in speech to target children was positively and significantly correlated with children's later vocabularies, whereas this link was not statistically significant when using model-derived predictions of adult speech was directed to others.

Suggested Uses of the Classifier

We constructed a web app (https://kachergis.shinyapps.io/classify_cds_ods/) deploying the final XGboost classifiers for both sleep and tCDS/ODS, so that other researchers with daylong LENA recordings can easily use it on their datasets. However, it is important to note that this app has only been trained with data from U.S. families; thus, for researchers with populations dissimilar to those studied here, we recommend checks for the reliability of the classifier against human listeners (see Limitations below). Additionally, research on the generalizability of the classifier to new samples deserves separate attention, especially when considering which variables are theoretically motivated and logistically possible under different circumstances.

For those with LENA data, use of this web app may facilitate specifying the amount of speech directed to target children and speech directed to others. First, the sleep classifier can automate one laborious step of 'cleaning' daylong LENA recordings with a reasonably high degree of reliability. Second, the tCDS/ODS classifier could also be used to reduce the significant hours of manual labor required for coding periods of target-child-directed or other-directed speech. We have found that the classifier's per-segment probability of tCDS matches well with the uncertainty of human coders (e.g., the 50/50 "split" segments were classified as ~50% probability of being tCDS).

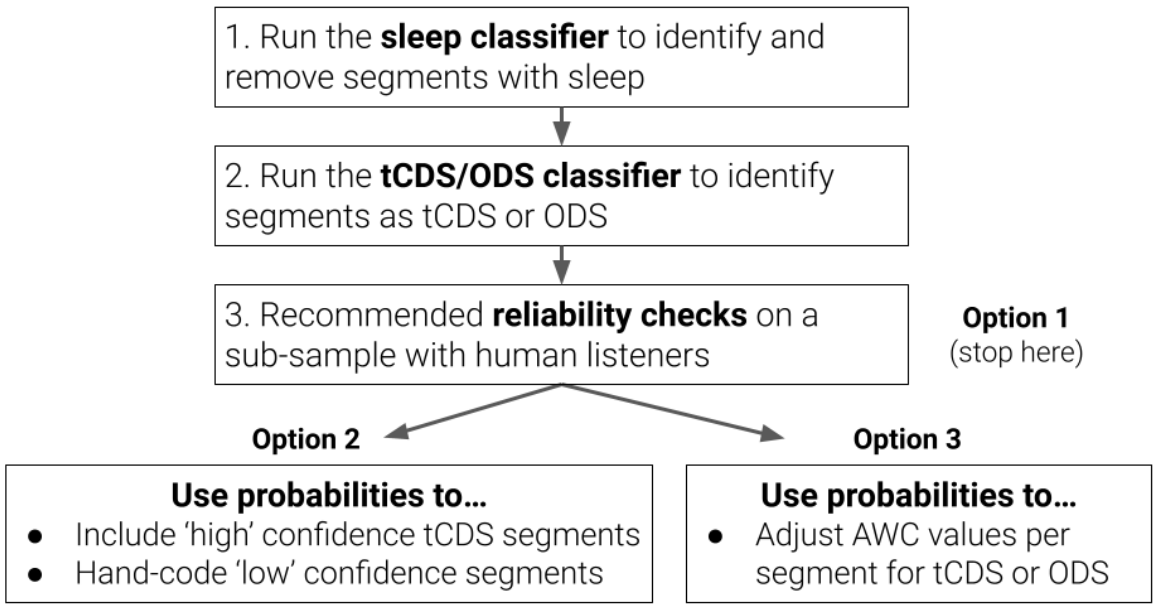
We suggest three potential workflows for using the classifier (Figure 9). Option 1 is to first run the sleep classifier to exclude periods when the child is sleeping and thus less likely to learn from the available speech, then running the tCDS/ODS classifier to identify binary judgements of segments considered as tCDS or ODS. Given that the classifier has been tested with a limited number of samples, we recommend reliability checks on a sub-sample of data with human listeners (if approved by ethics committees). To facilitate this, we provide our interrater reliability protocol for training human listeners, as well as the coding protocols from the original studies

(https://osf.io/qcj6r/). These protocols could also help guide reliability checks to compare human vs. classifier judgements.

Option 2 could be to follow the same steps, but rather than use a binary tCDS or ODS judgment, use the probabilities of tCDS or ODS to identify 'high confidence' tCDS or ODS segments versus 'low confidence' tCDS or ODS segments; 'low confidence' segments could then be listened to and judged by human coders. Whichever values are chosen, it is recommended to choose values that are symmetric (e.g., Pr(tCDS) < 0.3 (i.e., ODS) and Pr(tCDS) > 0.7 (i.e., tCDS)), to limit the introduction of bias.

Option 3 is to use the classifier probabilities to estimate the number of AWC tokens of tCDS and ODS in each segment by computing expected values (see Appendix D for more explanation). For example, a segment with an AWC of 200 and a .7 probability of being tCDS would result in 140 adult words counted as tCDS and 60 words counted as ODS for that segment. Rather than binning segments based on a binary probability of the entire segment falling into the tCDS versus ODS category, each 5-minute segment would contribute some of its counts to both. See Appendix D for an application of this method to the Weisleder & Fernald (2013) data, which yielded similar associations with outcomes. It is important to note that higher tCDS probabilities may reflect more of a certain type of verbal interaction (e.g., one-on-one interactions in a quiet indoor setting) than other types of caregiver-child interactions (e.g., playing outside where speakers may be further away from each other). Therefore, how probabilities are used should be considered with caution and transparently documented to better understand their utility and significance.

Figure 9. ***Potential workflows with the classifier.***



Limitations

While we included over 1,000 hours of data from 153 English- and Spanish-speaking families from varied socioeconomic backgrounds, our sample nevertheless represents a small subset of the variability that exists within English- and Spanish-speaking families in the U.S. and a tiny subset of the linguistic (e.g., different languages, multilingualism, signed vs. spoken language), cultural, and ecological variability in child-rearing environments around the world. For example, given the wide variability in infant sleep routines seen across families and countries throughout the world (Mindell et al., 2010), most of which are not represented in our training data, it is possible that the LENA features that characterize periods of sleep in our recordings will not generalize to recordings collected in very different contexts. In particular, all of the families in our studies lived in urban settings, and it is likely that the LENA features that characterize periods of sleep would differ for families in different settings (e.g., subsistence farming communities; Casillas et al., 2019, 2021). Similarly, given the wide variability in ways of interacting with children observed across sociocultural settings, it is possible that the features that differentiated tCDS from ODS in our sample of English- and Spanish-speaking families in California will not generalize to other contexts. Further validation studies are critical to understand whether our classifiers can generalize to new languages and communities (Cristia et al., 2021). Other studies that have coded tCDS vs. ODS in various other languages and contexts (Tseltal in a Mayan village: Casillas et al., 2019; Yélî Dnye in a Papuan community: Casillas et al., 2021; Spanish in Argentina: Rosemberg et al., 2020; Sesotho in South Africa and French in France: Loukatou et al., 2022) have done this in different ways (e.g., utterance-level coding vs. global binary judgements of tCDS or ODS). Thus, at the moment, our classifier cannot be applied to these data. Additionally, while our classifier is open-source, LENA software is not; thus, the ability to use this classifier requires a substantial cost to purchase the LENA recorders and software. Future work should compare whether our classifiers can be used with open-source speech algorithms (e.g., ALICE; Räsänen et al., 2021) to achieve similar performance. Finally, while the classifier can facilitate identification of periods of sleep, tCDS, and ODS in daylong audio recordings, this automated method does not reveal the specific acoustic, linguistic, or interactional features that differentiate between these speech contexts. Thus, it is far from replacing the need for human annotation and transcription and more research is needed to better explain how children learn from the language(s) to which they are exposed.

Conclusion

These findings suggest exciting opportunities for advancing our understanding of how children learn from the available speech in their environment. We were able to train and validate two automated classifiers using LENA-based measures to identify periods of sleep and to distinguish between periods of tCDS versus ODS. This work has the potential to significantly reduce the time-consuming process of identifying periods of directed speech to target children from the rich and naturalistic information collected with daylong recordings. In this way, the progress that we have

made here can facilitate future research seeking to illuminate questions about the relations between target-child-directed and other-directed speech on child outcomes and/or about the features of child-directed speech across linguistically- and culturally-diverse communities. We hope this adds to existing methods to explore shared and different features of target-child- and other-directed speech so we can better understand how different children across diverse communities acquire and develop their language skills.

References

Akhtar, N., Jipson, J., & Callanan, M. (2001). Learning words through overhearing. *Child Development, 72*(2), 416–430. https://doi.org/10.1016/S0163-6383(98)91471-0

Bakeman, R. (2023). KappaAcc: A program for assessing the adequacy of kappa. *Behavior Research Methods, 55*(2), 633-638. https://doi.org/10.3758/s13428-022-01836-1

Bang, J. Y., Mora, A., Munévar, M., Fernald, A., & Marchman, V. A. (2022). *Time to talk: Multiple sources of variability in caregiver verbal engagement during everyday activities in English- and Spanish-speaking families in the U.S*. PsyArXiv. https://doi.org/10.31234/osf.io/6jzwg

Busch, T., Sangen, A., Vanpoucke, F., & Van Wieringen, A. (2018). Correlation and agreement between Language ENvironment Analysis (LENA™) and manual transcription for Dutch natural language recordings. *Behavior Research Methods, 50*(5), 1921–1932. https://doi.org/10.3758/s13428-017-0960-0

Byers-Heinlein, K., Tsui, A. S. M., Bergmann, C., Black, A. K., Brown, A., Carbajal, M. J., Durrant, S., Fennell, C. T., Fiévet, A.-C., Frank, M. C., Gampe, A., Gervain, J., Gonzalez-Gomez, N., Hamlin, J. K., Havron, N., Hernik, M., Kerr, S., Killam, H., Klassen, K., … Wermelinger, S. (2021). A multilab study of bilingual infants: Exploring the preference for infant-directed speech. *Advances in Methods and Practices in Psychological Science, 4*(1), 2515245920974622. https://doi.org/10.1177/2515245920974622

Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., & Thai-Van, H. (2016). Reliability of the Language ENvironment Analysis system (LENA™) in European French. *Behavior Research Methods, 48*(3), 1109–1124. https://doi.org/10.3758/s13428-015-0634-8

Casillas, M., Brown, P., & Levinson, S. C. (2019). Early language experience in a Tseltal Mayan village. *Child Development, 91*(5), 1819–1835. https://doi.org/10.1111/cdev.13349

Casillas, M., Brown, P., & Levinson, S. C. (2021). Early language experience in a Papuan community. *Journal of Child Language, 48*(4), 792–814. https://doi.org/10.1017/S0305000920000549

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., & He, T. (2023). *xgboost: EXtreme Gradient Boosting* (1.7.5.1). https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf

Crago, M. B., Allen, S. E. M., & Hough-Eyamie, W.P. (1997). Exploring innateness through cultural and linguistic variation. In M. Gopnik (Ed.), *The inheritance and innateness of grammars* (pp. 70–90). New York City, NY, USA: Oxford University Press.

Cristia, A., Bulgarelli, F., & Bergelson, E. (2020). Accuracy of the Language Environment Analysis System segmentation and metrics: A systematic review. *Journal of Speech, Language & Hearing Research, 63*(4), 1093–1105. https://doi.org/10.1044/2020_JSLHR-19-00017

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J., & Bergelson, E. (2021). A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods, 53*(2), 467–486. https://doi.org/10.3758/s13428-020-01393-5

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 13*(4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005

Cychosz, M., Romeo, R., Soderstrom, M., Scaff, C., Ganek, H., Cristia, A., Casillas, M., de Barbaro, K., Bang, J. Y., & Weisleder, A. (2020). Longform recordings of everyday life: Ethics for best practices. *Behavior Research Methods, 52,* 1951-1969. https://doi.org/10.3758/s13428-020-01365-9

Cychosz, M., Villanueva, A., & Weisleder, A. (2021). Efficient estimation of children's language exposure in two bilingual communities. *Journal of Speech, Language, and Hearing Research, 64*(10), 3843–3866. https://doi.org/10.31234/osf.io/dy6v2

Dailey, S., & Bergelson, E. (2022). Language input to infants of different socioeconomic statuses: A quantitative meta-analysis. *Developmental Science, 25*(3), e13192. https://doi.org/10.1111/desc.13192

De Palma, P., & VanDam, M. (2017). Using automatic speech processing to analyze fundamental frequency of child-directed speech stored in a very large audio corpus. *IEEE Proceedings of IFSA-SCIS,* 1–6. https://doi.org/10.1109/IFSA-SCIS.2017.8023224

Ferjan Ramírez, N., Hippe, D. S., Braverman, A., Weiss, Y., & Kuhl, P. K. (2023). A comparison of automatic and manual measures of turn-taking in monolingual and bilingual contexts. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02127-z

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science, 16*(2), 234–248. https://doi.org/10.1111/desc.12019

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language, 16*(3), 477–501. https://doi.org/10.1017/S0305000900010679

Gampe, A., Liebal, K., & Tomasello, M. (2012). Eighteen-month-olds learn novel words through overhearing. *First Language, 32*(3), 385–397. https://doi.org/10.1177/0142723711433584

Gilkerson, J., & Richards, J. A. (2020). *A guide to understanding the design and purpose of the LENA system*. LENA Foundation. https://www.lena.org/wp-content/uploads/2020/07/LTR-12_How_LENA_Works.pdf

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H. L., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology, 26*(2), 248–265. https://doi.org/10.1044/2016_AJSLP-15-0169

Gilkerson, J., Richards, J. A., Warren, S. F., Oller, D. K., Russo, R., & Vohr, B. (2018). Language experience in the second year of life and language outcomes in late childhood. *Pediatrics, 142*(4), e20174276. https://doi.org/10.1542/peds.2017-4276

Gilkerson, J., Zhang, Y., Xu, D., Richards, J. A., Xu, X., Jiang, F., Harnsberger, J., & Topping, K. (2015). Evaluating Language Environment Analysis System Performance for Chinese: A Pilot Study in Shanghai. *Journal of Speech, Language, and Hearing Research, 58*(2), 445–452. https://doi.org/10.1044/2015_JSLHR-L-14-0014

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*(5), 515–523. https://doi.org/10.1111/j.1467-9280.2008.02117.x

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., Simson, J., Knox, D., Glowacki, L., Alemu, E., Galbarczyk, A., Jasienska, G., Ross, C. T., Neff, M. B., Martin, A., Cirelli, L. K., Trehub, S. E., Song, J., Kim, M., … Mehr, S. A. (2020). Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour*, 1–12. https://doi.org/10.1101/2020.04.09.032995

Hoff, E., Burridge, A., Ribot, K. M., & Giguere, D. (2018). Language specificity in the relation of maternal education to Bilingual Children's vocabulary growth. *Developmental Psychology, 54*(6), 1011–1019. https://doi.org/10.1037/dev0000492

Jackson-Maldonado, D., Thal, D., J., & Fenson, L. (2003). *MacArthur Inventarios del Desarrollo de Habilidades Comunicativas: User's guide and technical manual*. Brookes Publishing.

Lehet, M., Arjmandi, M. K., Houston, D., & Dilley, L. (2021). Circumspection in using automated measures: Talker gender and addressee affect error rates for adult speech detection in the Language ENvironment Analysis (LENA) system. *Behavior Research Methods, 53*(1), 113–138. https://doi.org/10.3758/s13428-020-01419-y

Loukatou, G., Scaff, C., Demuth, K., Cristia, A., & Havron, N. (2022). Child-directed and overheard input from different speakers in two distinct cultures. *Journal of Child Language, 49*(6), 1173–1192. https://doi.org/10.1017/S0305000921000623

Marchman, V. A., Bermúdez, V. N., Bang, J. Y., & Fernald, A. (2020). Off to a good start: Early Spanish-language processing efficiency supports Spanish- and English-language outcomes at 4½ years in sequential bilinguals. *Developmental Science, 23*(6), e12973. https://doi.org/10.1111/desc.12973

McGillion, M. L., Herbert, J. S., Pine, J. M., Keren-Portnoy, T., Vihman, M. M., & Matthews, D. E. (2013). Supporting early vocabulary development: What sort of responsiveness matters? *IEEE Transactions on Autonomous Mental Development, 5*(3), 240–248. https://doi.org/10.1109/TAMD.2013.2275949

Mendoza, J. K., & Fausey, C. M. (2021). Quantifying everyday ecologies: Principles for manual annotation of many hours of infants' lives. *Frontiers in Psychology, 12*, 710636. https://doi.org/10.3389/fpsyg.2021.710636

Mindell, J. A., Sadeh, A., Wiegand, B., How, T. H., & Goh, D. Y. T. (2010). Cross-cultural differences in infant and toddler sleep. *Sleep Medicine, 11*(3), 274–280. https://doi.org/10.1016/j.sleep.2009.04.012

Ochs, E., & Schieffelin, B. (1984). Language acquisition and socialization: Three developmental stories and their implications. In R. Schweder, & R. Levine (Eds.) *Culture theory: Essays on mind, self, and emotion* (pp. 276–322). Cambridge University Press.

Ohn-Bar, E., & Trivedi, M. M. (2016). To boost or not to boost? On the limits of boosted trees for object detection. *2016 23rd International Conference on Pattern Recognition (ICPR)*, 3350–3355. https://doi.org/10.1109/ICPR.2016.7900151

Quigley, J., Nixon, E., & Lawson, S. (2019). Exploring the association of infant receptive language and pitch variability in fathers' infant-directed speech. *Journal of Child Language, 46*(04), 800–811. https://doi.org/10.1017/S0305000919000175

Räsänen, O., Kakouros, S., & Soderstrom, M. (2018). Is infant-directed speech interesting because it is surprising? – Linking properties of IDS to statistical learning and attention at the prosodic level. *Cognition, 178,* 193–206. https://doi.org/10.1016/j.cognition.2018.05.015

Räsänen, O., Seshadri, S., Lavechin, M., Cristia, A., & Casillas, M. (2021). ALICE: An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. *Behavior Research Methods, 53*(2), 818–835. https://doi.org/10.3758/s13428-020-01460-x

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. E. (2018). Beyond the 30-Million-Word Gap: Children's conversational exposure is associated with language-related brain function. *Psychological Science, 29*(5), 700–710. https://doi.org/10.1177/0956797617742725

Rosemberg, C. R., Alam, F., Audisio, C. P., Ramirez, M. L., Garber, L., & Migdalek, M. J. (2020). Nouns and verbs in the linguistic environment of Argentinian toddlers: Socioeconomic and context-related differences. *First Language, 40*(2), 192–217. https://doi.org/10.1177/0142723719901226

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., Warlaumont, A. S., Hidalgo, G., Schnieder, S., Heiser, C., Hohenhorst, W., Herzog, M., Schmitt, M., Qian, K., Zhang, Y., … Zafeiriou, S. (2017). The INTERSPEECH 2017 Computational paralinguistics challenge: Addressee, cold & snoring. *Interspeech 2017,* 3442–3446. https://doi.org/10.21437/Interspeech.2017-43

Schuster, S., Pancoast, S., Ganjoo, M., Frank, M. C., & Jurafsky, D. (2014). Speaker-independent detection of child-directed speech. *2014 IEEE Spoken Language Technology Workshop (SLT),* 366–371. https://doi.org/10.1109/SLT.2014.7078602

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental Science, 15*(5), 659–673. https://doi.org/10.1111/j.1467-7687.2012.01168.x

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy, 14*(6), 654–666. https://doi.org/10.1080/15250000903263973

Snow, C. E. (1977). Mothers' speech research: From input to interaction. In C. Snow, & C. A. Ferguson (Eds.), *Talking to children* (pp. 31–49). Cambridge University Press.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review, 27*(4), 501–532. https://doi.org/10.1016/j.dr.2007.06.002

Soderstrom, M., & Wittebolle, K. (2013). When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. *PLoS ONE, 8*(11). https://doi.org/10.1371/journal.pone.0080646

Solomon, O. (2011). Rethinking baby talk. In A. Duranti, E. Ochs, & B. B. Schieffelin (Eds.), *The Handbook of Language Socialization* (1st ed., pp. 121–149). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444342901.ch5

Stärk, K., Kidd, E., & Frost, R. L. A. (2022). Word segmentation cues in German child-directed speech: A corpus analysis. *Language and Speech, 65*(1), 3–27. https://doi.org/10.1177/0023830920979016

Tamis-LeMonda, C. S., Kuchirko, Y., & Song, L. (2014). Why is infant language learning facilitated by parental responsiveness? *Current Directions in Psychological Science, 23*(2), 121–126. https://doi.org/10.1177/0963721414522813

Tamis-LeMonda, C. S., Song, L., Leavell, A. S., Kahana-Kalman, R., & Yoshikawa, H. (2012). Ethnic differences in mother-infant language and gestural communications are associated with specific skills in infants: Mother-infant communications. *Developmental Science, 15*(3), 384–397. https://doi.org/10.1111/j.1467-7687.2012.01136.x

Tomasello, M. (1995). Joint attention as social cognition. In C. Moore, P. Dunham, J. Philip (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). Hillsdale, NJ: Erlbaum.

VanDam, M., & Silbert, N. H. (2016). Fidelity of automatic speech processing for adult and child talker classifications. *PLoS ONE, 11*(8), e0160588. https://doi.org/10.1371/journal.pone.0160588

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science, 24*(11), 2143–2152. https://doi.org/10.3399/096016407782317928

Xu, D., Yapanel, U., Gray, S., & Baer, C., T. (2009). *The LENA Language Environment Analysis System: The Interpreted Time Segments (ITS) File*. LENA Foundation. https://www.lena.org/wp-content/uploads/2016/07/LTR-04-2_ITS_File.pdf

Yurovsky, D. (2018). A communicative approach to early word learning. *New Ideas in Psychology, 50*, 73–79. https://doi.org/10.1016/j.newideapsych.2017.09.001

Data, code and materials availability statement

All data and analysis code for the app are available here: https://github.com/kachergis/classify_cds_ods. All analysis code for the current manuscript are available here: https://github.com/kachergis/tCDS_nap_classifier_paper. Our protocol to determine human to human interrater reliability are available here: https://osf.io/qcj6r/

Ethics statement

Ethics approval was obtained from the Stanford University Institutional Review Board. All families provided informed consent prior to their participation in the study.

Authorship and Contributorship Statement

All authors contributed to conceptualization of the present study. J.B. (logistic regressions and reliability) and G.K. (classifiers) contributed to analyses. J.B., A.W., and V.M. supervised and analyzed original coding of LENA data across the five samples. J.B. and G.K. contributed to the original draft preparation. All authors contributed to writing, reviewing, and editing of the final manuscript.

Acknowledgements

We are especially grateful to the families for their contribution to this research. We would also like to thank Anne Fernald for supporting the studies that enabled this work, the Language Learning Lab staff for their tireless work in hand-coding these data, and the members of the Language and Cognition lab at Stanford and the LangVIEW consortium for their thoughtful comments and suggestions. This work was supported by grants from the National Institutes of Health (R01 HD42235, HD092343, HD069150), the Schusterman Foundation, the W.K. Kellogg Foundation, the David and Lucile Packard Foundation, and the Bezos Family Foundation to Anne Fernald, the National Institutes of Health (2R01 HD069150) to Heidi Feldman, the National Institutes of Health (R21 DC018357) and a Elizabeth Munsterberg Koppitz Child Psychology Graduate Student Fellowship from the American Psychological Foundation to Adriana Weisleder, and a Postdoctoral Support Award from the Stanford Maternal and Child Health Research Institute to Janet Bang. This work is an extension of work published in the Proceedings of the 46th annual Boston University Conference on Language Development (Bang, Kachergis, Weisleder, & Marchman, 2022).

Appendices

Appendix A. Final Sleep Classifier

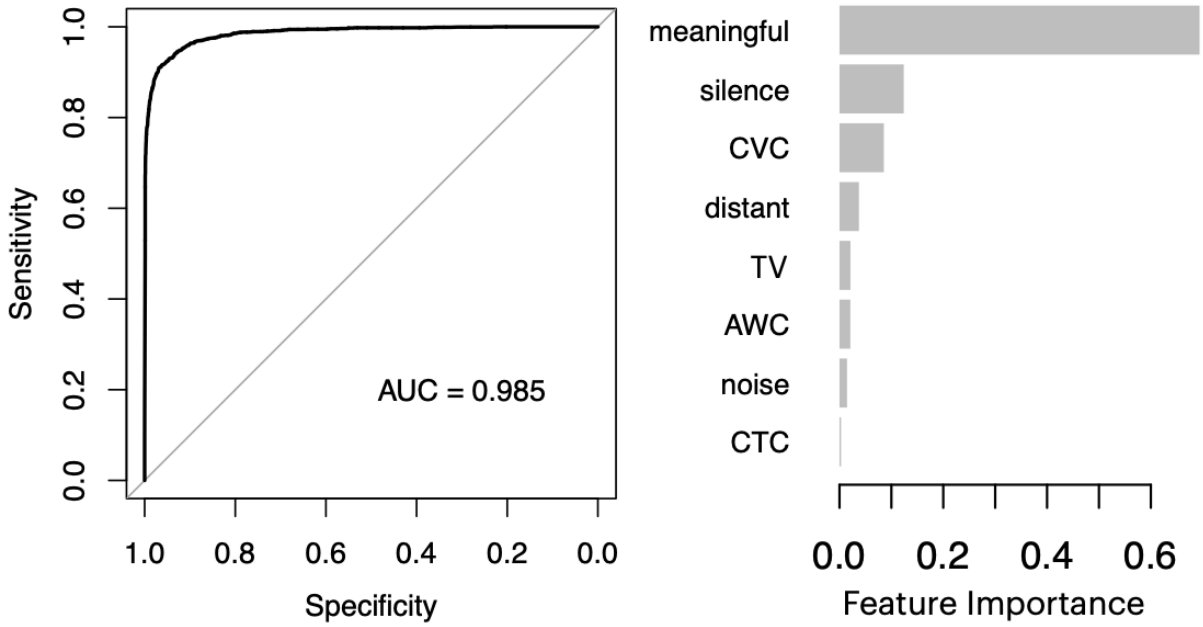


Figure A1. ***The final XGBoost sleep classifier, trained on the entire dataset, has a slightly higher AUC than the cross-validated classifier had on held-out data (Figure 5). The relative feature importances are quite similar to the held-out data classifier, although TV became slightly more important than AWC in the final classifier.***

Appendix B. Final tCDS/ODS Classifier

A final XGboost classifier was trained on the entire set of tCDS and ODS segments, and the results of this classifier are shown in Figure B1. The feature importances are similar to those in the classifier trained on 90% of the data, except that there is more reliance on AWC and slightly less on CTC in the final classifier. The AUC is also much improved.

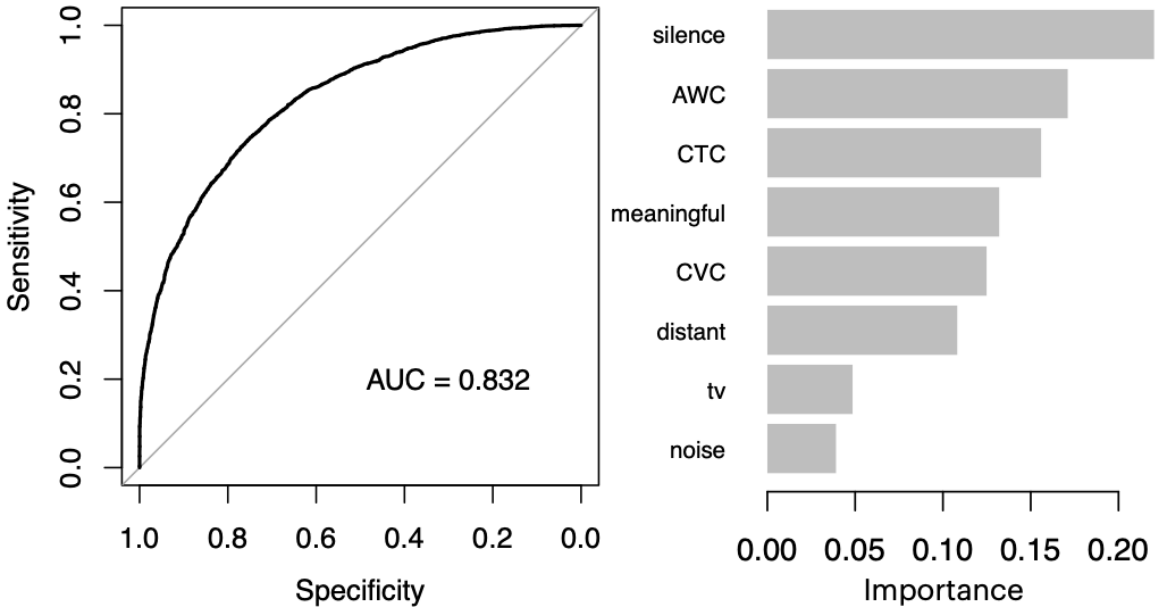


Figure B1. ***(left) ROC curve of the tCDS/ODS classifier and (right) relative importance of the LENA features in the final XGboost classifier trained on all 11,057 segments.***

Appendix C. Confusion Matrices Between Two Human Raters When Examining Interrater Reliability per Sample

Note that the diagonal (gray shading) indicates agreement between two human raters. For all tables, the percent agreement is calculated by dividing the number of agreements by the gold standard's total codes of the respective category.

Table C1. ***Sample 1***

		Human rater 2		
		tCDS	ODS	Total
Human Rater 1 (Gold Standard)	tCDS (rater 1)	46 (92% agreements)	4	50
	ODS (rater 1)	5	4	9

		(44% agreements)	
Total	51	8	59

Table C2. ***Sample 2***

		Human rater 2		
		tCDS	ODS	Total
Human rater 1 (Gold Standard)	tCDS (rater 1)	28 (100% agreements)	0	28
	ODS (rater 1)	11	17 (61% agreements)	28
	Total	39	17	56

Table C3. ***Sample 3***

		Human rater 2		
		tCDS	ODS	Total
Human rater 1 (Gold Standard)	tCDS (rater 1)	37 (84% agreements)	7	44
	ODS (rater 1)	9	25 (74% agreements)	34
	Total	46	32	78

Table C4. ***Sample 4***

		Human rater 2		
		tCDS	ODS	Total
Human rater 1 (Gold Standard)	tCDS (rater 1)	32 (76% agreements)	10	42
	ODS (rater 1)	1	21 (95% agreements)	22
	Total	33	31	64

Table C5. ***Sample 5***

		Human rater 2		
		tCDS	ODS	Total
Human rater 1 (Gold Standard)	tCDS (rater 1)	47 (87% agreements)	7	54
	ODS (rater 1)	13	7 (35% agreements)	20
	Total	60	14	74

Appendix D. Using Classifier Probabilities to Estimate tCDS and ODS Tokens

Future research may benefit from using the classifier-estimated probability of each segment being tCDS in two ways: 1) to reduce the amount of human coding (e.g, by only listening to the low-confidence segments), and 2) to estimate the number of tCDS and ODS tokens in each segment. First, given the high accuracy of the classifier for high-confidence classifications (~92% for Pr(tCDS) > 0.7 (i.e., tCDS), and 77% Pr(tCDS) < 0.3 (i.e., ODS)), one could use the classifier predictions for these segments, while potentially choosing to code the remaining low-confidence segments by hand. For the present dataset, this would have reduced the time needed to code the segments by 46%. For researchers primarily interested in segments that are likely to be primarily tCDS, it may be justified to disregard the likely ODS segments (e.g., Pr(tCDS) < 0.3; ~20% of our dataset). Determining what criterion to use requires careful consideration of the goals of the research, but there may be additional utility in leveraging the classifier's immediate, fine-grained judgments to support human rating for more difficult segments.

Moreover, the classifier's probability rating for each segment could be interpreted as an estimated proportion of the segment's tCDS (vs. ODS) content, and researchers could use the estimated tokens of tCDS and ODS AWC to calculate an expected value of both tCDS and ODS tokens for each child. That is, if a given 5-minute segment with 100 adult words receives a rating of Pr(tCDS) = 0.75, then the expected number of tCDS tokens in that segment is Exp(tCDS) = 100*0.75 = 75, and Exp(ODS) = 100*(1-Pr(tCDS)) = 25 tokens. Using this more fine-grained measure of each segment's contents may provide a better signal, as compared to the binarized classification, which assigns each segment's AWC tokens to either tCDS or ODS. Whether a segment with a higher probability of tCDS actually contains more tCDS (and less ODS) is an empirical question, which we will indirectly address here by examining the relation between children's classifier-rated amount of experienced tCDS and ODS and their later vocabulary size using the data from Weisleder & Fernald (2013), as before. The correlation for Exp(tCDS) and vocabulary size at 24 months was r = .56 (t(27) = 3.53, 95% CI = [.25, .77], p = .001), which is somewhat higher than when using the binary tCDS/ODS

judgments, from either the classifier or the human raters. The correlation for Exp(ODS) and vocabulary size at 24 months was r = .35 (t(27 = 1.94, 95% CI = [-.02, .64], p = .06), roughly similar to what was found using the binary judgments. Another hint that the classifier's Pr(tCDS) rating may correspond to humans' confidence is that the majority (74%) of the 'split' segments identified by human raters had great uncertainty for the classifier: only 26% of these segments were given high-confidence ratings in the model (Pr(tCDS) < 0.3 or Pr(tCDS) > 0.7).

Appendix E. Testing Classifiers Using Sample-Level Cross-Validation

Given that these samples were collected over many years, with potential variation in populations and training of research assistants, we chose to test whether leaving contemporaneously collected samples out of the training set unduly influenced the performance of the sleep or tCDS/ODS classifiers. Table E1 shows the accuracy and AUC for sleep classifiers trained without each sample, showing that performance was fairly consistent (accuracy range: [0.945,0.970]; AUC range: [0.948,0.985]). Table E2 shows the results for tCDS/ODS classifiers trained without each sample, showing broadly similar performance (accuracy range: [0.628,0.708]; AUC range: [0.690,0.786]). It is worth noting that leaving out Sample 1 does somewhat decrease performance, and leaving out Sample 4 somewhat increases performance. Nonetheless, on balance we believe that including the full dataset gives the greatest chance of generalizing to new datasets.

Table E1. ***Sleep classifier results when respective samples are left out***

Sample-left-out	Accuracy of classifier without sample	Area Under Curve (AUC) without sample
Sample 1	0.970	0.985
Sample 2	0.945	0.955
Sample 4	0.948	0.966
Samples 3 and 5*	0.967	0.948

Table E2. ***tCDS/ODS classifier results when respective samples are left out***

Sample-left-out	Accuracy of classifier without sample	Area Under Curve (AUC) without sample
Sample 1	0.628	0.690
Sample 2	0.666	0.721
Sample 4	0.708	0.786
Samples 3 and 5*	0.683	0.723

Note: *Samples 3 and 5 were both coded at the 10-min level, and then split into 5-min segments to include in the classifier. We group them here to cross-validate the classifier.

Appendix F. Testing Classifiers Using Child-Level Cross-Validation

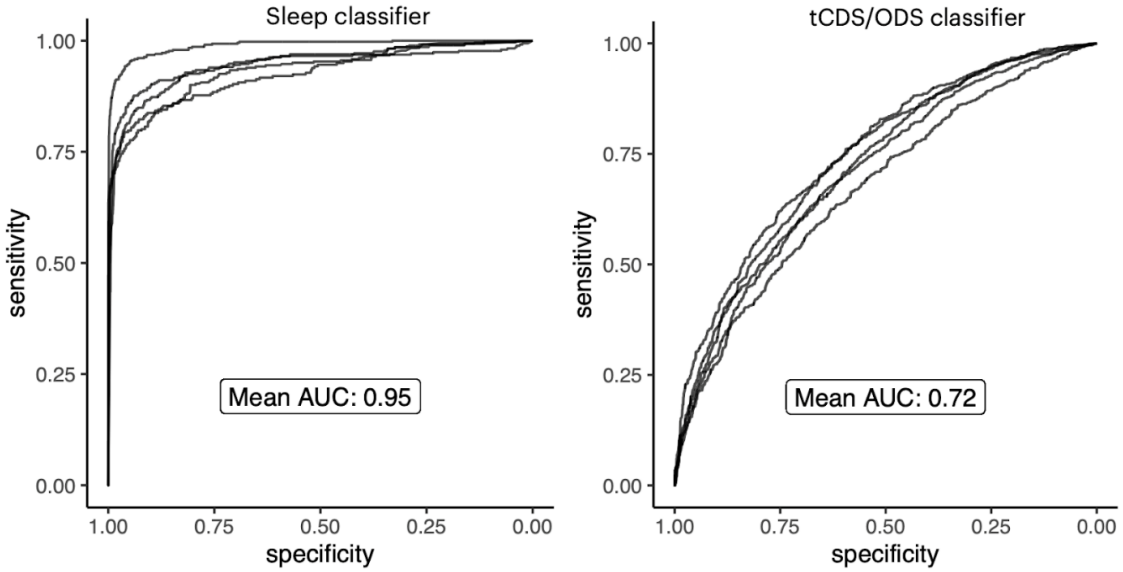


Figure F1. ***ROC curves for classifiers that exclude 20% of children (n = 30) in each training set, for sleep (left) and tCDS/ODS (right).***

License

Bilingual children's comprehension of code-switching at an uninformative adjective

Lena V. Kremin
Concordia University, CA

Amel Jardak
Concordia University, CA

Casey Lew-Williams
Princeton University, USA

Krista Byers-Heinlein
Concordia University, CA

Abstract: Bilingual children regularly hear sentences that contain words from both languages, also known as code-switching. Investigating how bilinguals process code-switching is important for understanding bilingual language acquisition, because young bilinguals have been shown to experience processing costs and reduced comprehension when encountering code-switched nouns. Studies have yet to investigate if processing costs are present when children encounter code-switches at other parts of speech within a sentence. The current study examined how 30 young bilinguals (age range: 37 – 48 months) processed sentences with code-switches at an uninformative determiner-adjective pair before the target noun (e.g., "Can you find *le bon* [the good] duck?) compared to single-language sentences (e.g., "Can you find the good duck?"). Surprisingly, bilingual children accurately identified the target object in both sentence types, contrasting with previous findings that sentences containing code-switching lead to processing difficulties. Indeed, children showed similar (and in some cases, better) comprehension of sentences with a code-switch at an uninformative adjective phrase, relative to single-language sentences. We conclude that functional information conveyed by a code-switch may contribute to bilingual children's sentence processing.

Keywords: code-switching; bilingualism; language processing; language acquisition

Corresponding author(s): Krista Byers-Heinlein, Department of Psychology, Concordia University, 7141 Sherbrooke St. West, Montreal, QC, H4B 1R6. Email: k.byers@concordia.ca.

ORCID ID(s): Lena V. Kremin: https://orcid.org/0000-0002-4010-490X
Casey Lew-Williams: https://orcid.org/0000-0002-8781-4458
Krista Byers-Heinlein: https://orcid.org/0000-0002-7040-2510

Citation: Kremin, L.V., Jardak, A., Lew-Williams, C., & Byers-Heinlein, K. (2023). Bilingual children's comprehension of code-switching at an uninformative adjective. *Language Development Research*, *3*(1), 249–276. https://doi.org/10.34842/zyvj-cv60

Introduction

Bilingual children regularly hear both of their languages within a single conversation and even within a single sentence (e.g., *C'est un* [fr. It's a] monkey.). This phenomenon is known as code-switching. Most bilingual children hear code-switching in their daily lives (Kremin et al., 2021), and there is some evidence that over time code-switching may impact a child's vocabulary size (Bail et al., 2015; Byers-Heinlein, 2013) and overall language development (Kaushanskaya & Crespo, 2019). Code-switching can also reduce a child's comprehension in the moment as they process speech (Byers-Heinlein et al., 2017; Morini & Newman, 2019; Potter et al., 2019). Given the importance of early language processing to early language acquisition (Meylan & Bergelson, 2022), it is important to understand the contexts in which code-switching does or does not affect language processing, as well as underlying mechanisms. To date, research on children's comprehension of code-switching has focused on code-switches at a noun (e.g., "*Dónde está la* [sp. where's the] ball?"), even though everyday code-switching happens at many different parts of speech, such as verbs, prepositions, and adjectives (e.g., "*C'est* [fr. It is] yucky."; MacSwan, 2012). Here, we extend previous findings with nouns and investigate how code-switching at a mid-sentence determiner-adjective pair affects bilingual children's language comprehension.

A large body of literature has reported that bilingual adults process code-switches more slowly than single-language stimuli (for recent reviews see Beatty-Martínez et al., 2018; Valdés Kroff et al., 2018; van Hell et al., 2018), but researchers have only recently begun to study how young children process code-switches. One eye-tracking study indicated that children process code-switches differently depending on whether the switch happens between sentences or within a single sentence. When hearing between-sentence code-switching (e.g., "That one looks fun! *Le chien* [fr. the dog]!"), 1.5- to 2-year-old children were as accurate at identifying the target object as they were when hearing a single language (e.g., "That one looks fun! The dog!"; Byers-Heinlein et al., 2017). However, when hearing within-sentence code-switching (e.g., "Look! Find the *chien* [fr. dog]!"), children were less accurate at identifying the target object compared to hearing a single language (e.g., "Look! Find the dog!"; Byers-Heinlein et al., 2017; Morini & Newman, 2019). Such studies with young children have focused solely on code-switches at the noun, so they do not address the potential impact of code-switching at other parts of speech. This limitation makes it impossible to draw generalized conclusions about how code-switching may or may not affect comprehension. Children may process code-switching at different parts of speech more readily depending on several factors, such as how often children hear code-switching in that location or what functional information is contained in the code-switched word(s). Evaluating children's comprehension of code-switching at different parts of speech can provide insights into the veracity of two general accounts of what makes

code-switching difficult to process, which we introduce here as the frequency account and the functional account.

Frequency Account

The frequency account posits that how easily bilinguals process a code-switch depends on how frequently that type of code-switched construction occurs in their everyday life (e.g., Abutalebi et al., 2007; Guzzardo Tamargo et al., 2016; Salig et al., 2023). This account predicts that frequent code-switched constructions will be more easily processed than infrequent code-switched constructions. For example, in one study, Spanish–English bilingual adults more readily processed a common code-switch that included an entire compound verb (e.g., "*los senadores* [sp. the senators] **have requested** the funds") than an uncommon code-switch that occurred in the middle of the compound verb (e.g., "*los senadores* ***han*** [sp. the senators **have**] **requested** the funds"; Valdés Kroff et al., 2018). Similarly, Welsh–English bilingual adults judged code-switching at common parts of speech, such as nouns, to be more acceptable than code-switching at uncommon parts of speech, such as adjectives (Vaughan-Evans et al., 2020). The frequency account could also predict differences in comprehension between bilingual populations if they hear different rates of code-switching in their daily lives (Gosselin & Sabourin, 2021; Valdés Kroff et al., 2018).

If frequency is indeed an important factor in how bilingual adults process code-switching, its importance could also extend to children's processing. Under the frequency account, children would be expected to understand code-switching at frequently code-switched parts of speech, such as nouns, more easily than at infrequently code-switched parts of speech, such as adjectives. This account could explain existing findings about children's processing of code-switching. Indeed, when children do hear within-sentence code-switches, they often occur at nouns (Bail et al., 2015). Moreover, children hear more between-sentence code-switches than within-sentence code-switches from their parents (Bail et al., 2015; Kremin et al., 2021), so the frequency account is consistent with the experimental finding that children more easily process between-sentence code-switches compared to within-sentence code-switches (Byers-Heinlein et al., 2017; Morini & Newman, 2019). Thus, if within-sentence code-switches at a relatively common location for code-switching (i.e., the noun) can disrupt children's processing, then within-sentence code-switches at an uncommon location should be even more disruptive.

Functional Account

Another account – related to yet different from the frequency account – proposes that bilinguals process code-switches differently based on the *functional* properties of the code-switched word(s), including grammatical properties. While prior research has investigated a variety of functions of code-switching in production – such as adding

emphasis, signaling community identity, and facilitating understanding (Goodz, 1989; Heredia & Altarriba, 2001; Nilep, 2006) – comprehension studies have mainly focused on the functional dimension of grammatical class. One study of German–Russian bilingual adults used event-related potentials (ERPs) to examine the processing of code-switches at open-class words (e.g., nouns) versus closed-class words (e.g., prepositions). While code-switches at both nouns and prepositions elicited a broad late positivity, only code-switches at prepositions elicited a broad early negativity, suggesting that bilinguals process code-switches differently based on their grammatical function (Zeller, 2020). Another ERP study compared how bilinguals processed code-switching at two types of open-class words: nouns and verbs (Ng et al., 2014). When reading a story, Spanish–English bilingual adults processed code-switching at nouns (e.g., "the wind and the *sol* [sp. sun]") differently than code-switching at verbs (e.g., "they *miraron* [sp. saw] a traveler") as indicated by larger N400 responses and an early Late Positive Component for nouns. The authors proposed that the difference was driven by the effort bilinguals put into integrating and remembering the information contained in each code-switch. That is, nouns are likely to be referenced several times in a story and need to be held in working memory, thus eliciting more cognitive effort compared to verbs that may only be used once. Combined, these results highlight that bilinguals may be sensitive to the functional role of the code-switched words and process them accordingly.

Research has yet to investigate how bilingual children process code-switches with diverse functional or grammatical roles, but evidence from monolinguals shows that children are sensitive to some grammatical classes beginning around 8 months of age (Marino et al., 2020). Moreover, by age 3, children use the meaning of adjectives to predict which noun they refer to (e.g., predicting "heavy" is more likely to be followed by "stone" than "butterfly"; Tribushinina & Mak, 2016). Additionally, monolingual children as young as 2 years old can recognize, but "listen through," uninformative adjectives to quickly and correctly identify a target noun (Thorpe & Fernald, 2006). For example, when shown a picture of a dog and a bunny, children identified the target object as quickly when it was preceded by an uninformative adjective [e.g., "Where's the good bunny?") as when it was not preceded by any adjective (e.g., "Where's the bunny?"). These results show that young children can attend to the most relevant functional information to efficiently process speech.

Following the functional account, code-switching that occurs at a word that is central to the meaning of the sentence may be particularly challenging for children to process. In many cases, this will be a noun, but in other cases it could be a verb, adjective, or other part of speech, depending on context. This idea is supported by previous research showing that children experience difficulty in understanding functionally-important code-switched nouns (Byers-Heinlein et al., 2017; Morini & Newman, 2019). In contrast, code-switches at parts of speech that play a limited functional role in

comprehension may be relatively easy for children to process, and code-switches that are uninformative in a comprehension task may not elicit any processing difficulties. However, to date, children's comprehension of code-switches at words with limited functional meaning has not yet been investigated; thus there is a lack of empirical evidence for the functional account with children.

Current Study

In the current study, we asked if code-switching within a sentence at an uninformative determiner-adjective pair (which we will hereafter refer to as an uninformative adjective) affects children's comprehension of a target noun that immediately follows it. This allowed us to examine the potential contributions of frequency and/or functional factors in children's processing of code-switching. The frequency account predicts that children will show disrupted processing of a code-switch at an adjective, because it is not a common location for code-switching. This could result in weaker comprehension of the following noun, as processing difficulties earlier in the sentence can negatively affect how children process the end of the same sentence (Trueswell et al., 1999). In contrast, the functional account predicts that children may find it relatively easy to process a code-switch at an uninformative adjective as they do not necessarily have to attend to or remember its meaning in the context of the visual scene.

In an eye-tracking experiment, children viewed pairs of pictures of animals, such as a duck and a fish, and heard sentences such as "Can you find *le bon* [fr. the good] duck?" or "Can you see *el buen* [sp. the good] duck?" In trials, both animals were equally consistent with the adjective (e.g., both were depicted as equally "good"). Participants were 30 3-year-old bilinguals, including both French–English bilingual children in Montreal (n = 19) and Spanish–English children in New Jersey (n = 11). This age group was chosen because, from this age, children can attend to the information in adjectives in real-time sentence comprehension (Tribushinina & Mak, 2016). We included participants from these two testing locations to increase sample size, as bilingual children are a difficult-to-recruit population. This is in line with various sampling strategies in the field of early bilingualism which range from testing homogeneous populations (e.g., all acquiring English and French) to testing heterogeneous populations (e.g., all acquiring English and a variety of other languages; Byers-Heinlein, 2015). Assessing the effects of code-switching at adjectives was appropriate in our sample, because children of this age can generally understand their meaning (Tribushinina & Mak, 2016), and because certain adjectives can occur in the same prenominal position across the languages being acquired by our participants (i.e., English, French, and Spanish).

Similar to previous studies on children's processing of code-switching (Byers-Heinlein et al., 2017; Morini & Newman, 2019; Potter et al., 2019), we expected that code-switching at an uninformative adjective would hinder children's comprehension of the target noun compared to sentences without code-switching. Specifically, we predicted that children would look less towards the target noun after hearing mid-sentence code-switching compared to hearing a sentence entirely in one language. Such a result would be consistent with the frequency account. In contrast, a finding that children's performance was unaffected by an uninformative code-switched adjective would be consistent with the functional account. We also explored whether individual differences such as language dominance, testing location (as a proxy for language pair), SES, or vocabulary size would be related to performance.

Method

Data collection occurred in two locations: Montreal, Canada and New Jersey, USA. The methods were approved by the Concordia University Human Research Ethics Committee ("Monolingual and Bilingual Language Development"; approval #10000493) and the Princeton University Institutional Review Board ("Language learning and Communication"; approval #7117), and parents provided informed consent prior to their child's participation. Data were collected in Montreal between November 2016 and April 2017 and in New Jersey between March 2017 and January 2018. Final data analysis occurred between May 2020 and June 2021, during the COVID-19 pandemic. As is common in laboratories testing hard-to-recruit populations such as bilingual children, children participated in a second, separate study, either immediately prior to or following participation in this study (the order of the two studies was counterbalanced). The results of that study are reported in a separate manuscript (Byers-Heinlein et al., 2021). All stimuli, data, and analysis scripts for the current study are available via the Open Science Framework at https://osf.io/ecqwr/.

Participants

A total of 30 3-year-old (M = 3.57, range = 3.10 – 4.05, 14 females) full-term, healthy bilingual children participated in this study. This sample size was sufficiently sensitive to detect an effect size of d = 0.46 at 80% power in a paired-samples t-test, meaning there were enough participants to detect effect sizes reported in previous related studies (0.56 in Byers-Heinlein et al. 2017; 0.60 in Potter et al. 2019).

Nineteen French–English bilinguals were tested in Montreal, Canada, and 11 Spanish–English bilinguals were tested in New Jersey, USA. In Montreal, children were recruited from a database of families interested in participating in our research, principally identified via government birth lists. In New Jersey, children were primarily recruited from nonprofit organizations. Another 34 children were tested but not

included in the final sample due to not meeting the language criteria (n = 15; see details below), fussiness or lack of attention (n = 10), technical issues (n = 4), health reasons such as low birth weight or gestation period under 37 weeks (n = 3), completing an insufficient number of trials (n = 1; see below), or having a reported speech delay (n = 1). Post-hoc data exclusion resulted in the unbalanced sample between the two locations. Unfortunately, because this discrepancy did not become clear until the time of data analysis, which occurred during the COVID-19 pandemic, we were unable to test additional participants to address this difference. Parents reported their child's ethnicity/race using categories and a free-response option appropriate to each location. Among French-English bilinguals in Montreal, 10 children were European, 2 were Canadian, 2 were Caribbean, 1 was Arab, 1 was Quebecois, and 3 did not report. Among Spanish-English bilinguals in New Jersey, 8 children were Hispanic, 1 was Black, 1 was White, and 1 was from multiple ethnic/racial backgrounds.

Language Background and Proficiency

Children's language background and proficiency was assessed via a modified version of the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007). Parents were asked about their child's experience with each language, and to rate their child's proficiency in English and French (in Montreal) or in English and Spanish (in New Jersey) compared to monolingual children of the same age. Following a pre-determined inclusion criterion, children had to receive a comprehension score of at least 7/10 for both languages to be eligible for the study, to ensure that children were reasonably proficient in both languages. For each child, their dominant language was established as the language that had the highest comprehension score from the LEAP-Q. Twelve children had equal comprehension scores in both languages, so for these children, the language in which the child had the higher productive vocabulary score (see below) was considered their dominant language. In total, 19 children were dominant in English, 9 were dominant in French, and 2 were dominant in Spanish. Twelve children were regularly exposed to both of their languages from birth, and 18 children were exposed to their second language later in life, between the ages of 2 and 36 months. See Table 1 for details by testing location.

Vocabulary Size

Children's productive vocabulary size in English was assessed using the Developmental Vocabulary Assessment for Parents (DVAP; Libertus et al., 2015), which consisted of a checklist of words known by children aged 2 to 18 years old based on words used in the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 2007). We used a parent checklist rather than a direct measure to reduce children's fatigue, as each child participated in two experiments, and we wished to assess their vocabulary in both languages. Moreover, the DVAP has shown strong convergent validity with children's performance on the PPVT (β = .69; Libertus et al., 2015).

Table 1. ***Demographics of participants at each testing location.***

Location	*n*	Mean age in years (range)	Eng. dom. (*n*)	L2 from birth (*n*)	Later L2 (age range in months)	Dom. Lang. Vocab. (*SD*)	Non-Dom. Lang. Vocab. (*SD*)	Parent education (*SD*)
Montreal	19	3.47 (3.1 – 3.99)	10	8	6 – 18	76.83 (33.91)	47.83 (30.19)	16.58 (2.17)
New Jersey	11	3.75 (3.19 – 4.05)	9	4	2 – 36	62.36 (26.22)	24.55 (18.34)	12.82 (5.06)

Note. Eng. dom. (*n*) lists the number of children at each testing location who were dominant in English; the remainder of children were dominant in either French if tested in Montreal or Spanish if tested in New Jersey. Later L2 (age range in months) only considers participants who were not exposed to both languages from birth.

To assess children's productive vocabulary size in French or Spanish, we adapted a checklist similar to the DVAP, based on words used in the adaptation of the PPVT for Quebec French (Échelle de Vocabulaire en Images Peabody; Dunn et al., 1993) or Spanish (Test de Vocabulario en Imagenes Peabody; Dunn et al., 1986). The words are ordered from easy (e.g., "ball", "dog") to hard (e.g., "honing", "angler"), and parents were asked to indicate which words their child could say. There are 212 items on the English version, 190 items on the French version, and 125 items on the Spanish version. A parent or other adult that was familiar with the child's vocabulary in a particular language filled out the form for that language. In some cases, the forms for each language were completed by different parents who normally interacted with their child in that language, while in other cases it was one parent who filled out both forms if they used both languages with their child. As expected, the number of words children produced in their dominant language (M = 71, SD = 32, range = 24 – 177) was greater than the number of words they produced in their non-dominant language (M = 39, SD = 28, range = 2 – 131), $t(28) = 7.03$, $p < .001$, $M_D = 32.34$, 95% CI [22.92,41.77]). When combining the number of words produced in both languages, on average, children produced 110 total words (SD = 55, range = 31 – 308). Children in Montreal (M = 125, SD = 61, range = 39 – 308) produced more words than those in New Jersey (M = 87, SD = 33, range = 31 – 138), $t(26.73) = -2.16$, $p = .040$, $\Delta M = -37.76$, 95% CI [−73.56, −1.95]), although we note that there were more words on the French version than on the Spanish version of the DVAP.

Exposure to Parental Code-Switching

Children's exposure to parental code switching was measured via the Language Mixing Scale (Byers-Heinlein, 2013), which measures intra-sentential code switching by the primary caregiver. Scores can range from 0 (no switching) to 30 (highest amount of switching). In Montreal, caregivers' average score was 13.50, range = 4 – 28. In New Jersey, caregivers' average score was 14.80, range = 0 – 30. The difference in the amount of switching between the two locations was not significant, $\boldsymbol{t(15.88) = -0.39}$, $\boldsymbol{p = .700}$, $\boldsymbol{\Delta M = -1.34}$, 95% CI $\boldsymbol{[-8.61, 5.92]}$.

Socioeconomic Status

As a proxy for socioeconomic status (SES), we asked parents to indicate the highest level of education they had attained. As the education systems are somewhat different in the United States and Canada, to be able to compare responses across our two testing locations, we converted these responses to the typical number of years after kindergarten to complete each level of education (e.g., completing a bachelor's degree was equivalent to 16 years of education). For families where both parents' education was provided, the higher level was selected for analysis. On average, parents completed 15.20 (SD = 3.89) years of education, which ranged widely from 4 to 21 years. Parents in Montreal reported completing more years of education (M = 16.58, SD = 2.17, range = 13 – 21) than parents in New Jersey (M = 12.82, SD = 5.06, range = 4 – 20), $t(12.17) = 2.35$, $p = .037$, $\Delta M = 3.76$, 95% CI [0.27,7.25], suggesting that the participants in Montreal came from a higher SES background than those in New Jersey.

Material

Visual Stimuli

Visual stimuli consisted of 8 pairs of pictures for each language combination (See Table 2 for picture pairs and Figure 1 for an example trial). Each picture in a pair had the same animacy status (i.e., four pairs of animals used in target trials and four pairs of inanimate objects used in filler trials), so that the two pictures had similar visual salience. To ensure that they would be familiar to our 3-year-old participants, we selected pictures whose labels were highly understood by children in American English (Fenson et al., 2007), Quebec French (Boudreault et al., 2007), and Spanish (Jackson-Maldonado et al., 2003). The labels of the picture pairs did not overlap in word onset, had the same grammatical gender in French or Spanish, and are widely used across French and Spanish dialects. Pictures were chosen from free online libraries and digitally edited as necessary.

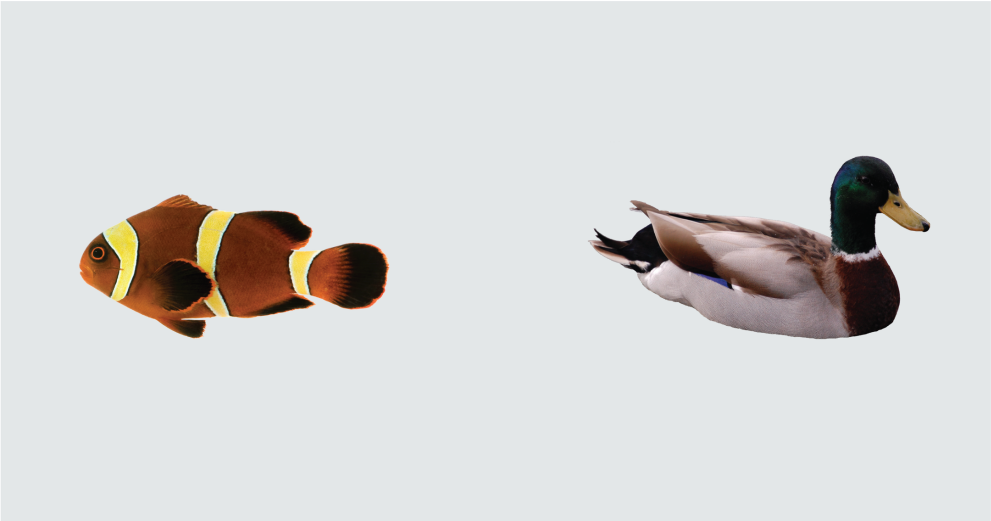


Figure 1. ***Example and timeline of experimental trial.***

Auditory Stimuli

Auditory stimuli were recorded by a female, native French–English or Spanish–English bilingual with no perceptible accent in either language using infant-directed speech. Each auditory stimulus contained a target word labeling one of the pictures on the screen (e.g., "Look! Can you find the good duck?"). The target noun (e.g., "duck") was preceded by a determiner (e.g., "the") and a prenominal adjective (e.g., "good"). Each stimulus sentence was recorded in a single-language version where the determiner and adjective were in the same language as the noun, and a code-switched version where the determiner and adjective were in the other language (e.g., "Look! Can you find *le bon* [fr. the good] duck?" or "Look! Can you see *el buen* [sp. the good] duck"?). Note that the target word (e.g., "duck") was always in the same language as the initial carrier phrase (e.g., "Look! Can you find…" for French–English and "Look! Can you see…" for Spanish–English). Parallel stimulus sets were created with the carrier sentences in each language (e.g., in French, the previous examples became "*Regarde! Peux-tu trouver le bon canard*?" and "*Regarde! Peux-tu trouver* the good *canard*?"; in Spanish, the previous examples became "*¡Mira! Puedes ver el buen pato*?" and "*¡Mira! Puedes ver* the good *pato*?").

For the animate nouns on target trials, there were a total of four English prenominal adjectives and their French and Spanish translations; similarly, for inanimate nouns in filler trials, there were four prenominal adjectives used (see Table 2). These adjectives were chosen such that they 1) were not cognates across French and English or Spanish and English, 2) did not share phonological overlap with their translation, 3) were not descriptive of one picture more than another, and 4) could precede a noun in French or Spanish. Although both French and Spanish usually place adjectives in a postnominal position, the adjectives we selected can be used prenominally in these grammatical contexts. Each adjective was always used with the same picture pair. All stimuli are available at https://osf.io/ecqwr/.

Trial Description

During each trial, the target and distractor pictures appeared on the screen for 6000ms, and one of the stimulus sentences was played labeling the target picture. The onset of the target noun occurred exactly 3000ms into each trial. The determiner–adjective combinations were of somewhat different lengths, and so occurred between 311 and 1152ms before the noun onset. Trials were combined into four experimental orders of 24 trials: 8 single-language trials (e.g., "Look! Can you find the good duck?"), 8 code-switched trials (e.g., "Look! Can you find le bon [fr. the good] duck?"), and 8 additional single-language filler trials. Filler trials were not analyzed and were mainly used to lower the overall number of trials with code-switching. Target trials (i.e., single-language and code-switched trials) and filler trials were intermixed throughout the study. The language of the carrier phrase was consistent for each child (i.e., always in English, French, or Spanish), but counterbalanced across children at the time of testing. In total, 15 children were tested with carrier phrases in their dominant language (10 French–English and 5 Spanish–English), and 15 children were tested with carrier phrases in their non-dominant language (9 French–English and 6 Spanish–English).

Procedure

In addition to signing a consent form, parents completed questionnaires on their child's vocabulary (DVAP) and language comprehension (LEAP-Q), on their own language mixing (Byers-Heinlein, 2013), and on basic demographic information. During the study, parents listened to music with headphones, wore darkened glasses, and were instructed not to interfere with the study or provide their child with any instruction. Testing occurred in a darkened room while children sat on their parent's lap.

Table 2: ***Adjective–noun pairs used for French–English and Spanish–English participants. The noun pairs labeled the two pictures shown on screen at the same time. Each noun was used as a target picture in different trials. In single-language trials, the adjective and noun were in the same language. In code-switched trials, the adjective and the noun were in different languages.***

English		**French**	
Look! Can you find … ?		*Regarde! Peux-tu trouver … ?*	
Adjective	**Noun pair**	**Adjective**	**Noun pair**
Target trials			
the good	duck – fish	le bon	canard – poisson
the little	monkey – sheep	le petit	singe – mouton
the nice	dog – bunny	le gentil	chien – lapin
the pretty	cow – froggy	la jolie	vache – grenouille
Filler trials			
a large	ear – spoon	une grosse	oreille – cuillère
a new	apple – toothbrush	une nouvelle	pomme – brosse à dents
a big	door – hand	une grande	porte – main
an old	coat – pencil	un ancien	manteau – crayon

English		**Spanish**	
Look! Can you see … ?		*¡Mira! ¿Puedes ver … ?*	
Adjective	**Noun pair**	**Adjective**	**Noun pair**
Target trials			
the good	bear – duck	el buen	oso – pato
the little	butterfly – sheep	la pequeña	mariposa – oveja
the big	bunny – dog	el gran	conejo – perro
the pretty	cow – froggy	la hermosa	vaca – rana
Filler trials			
a beautiful	ear – spoon	una linda	oreja – cuchara
a new	apple – toothbrush	una nueva	manzana – cepillo de dientes
a nice	door – hand	una preciosa	puerta – mano
an old	coat – pencil	un viejo	chamarra – lápiz

Due to differences in lab equipment, the same apparatus was not available at both testing sites. In Montreal, the study was conducted in the lab on a 24-inch Tobii T60XL corneal reflection eye-tracking system using a 5-point calibration, with auditory stimuli played over speakers. In New Jersey, the study was conducted either in the lab (7 children) or at a local community center (4 children), depending on which location was easier for participants to access. In the lab, visual and auditory stimuli were presented using Matlab on a 55" TV monitor. At the community center, visual stimuli were presented in a QuickTime video on a 13" laptop, and auditory stimuli were played through noise-canceling headphones. In both New Jersey setups, a video camera below the screen recorded children's eye movements at a rate of 30 frames per second for later offline coding by trained research assistants.

Before each trial began, a colorful attention-getter was presented to draw the child's attention to the screen. Once the child was looking at the screen, the trial began. An experimenter monitored the status of the study via video camera and controlled the experiment from a computer in another room (Montreal) or within the same room (New Jersey). The total duration of the study was approximately 4 minutes.

Coding

In Montreal, the eye-tracking system collected data on the location of children's eye-gaze and their pupil size at a rate of 60Hz. We defined areas of interest corresponding to a rectangle of 2 cm around each picture presented on the screen. In New Jersey, a trained research assistant used EyeCoder software to code at 33-ms intervals whether the child was looking at the left or right object on the screen, shifting between objects, or inattentive. A second research assistant coded 18% of videos; on the frames surrounding eye movements, inter-coder reliability was 97%. Research suggests that automatic eyet racking and manual gaze coding, although potentially different in their amount of data loss, capture largely similar information (Venker et al., 2020). We did not observe a difference in data loss between the two coding methods. An average of 15.88% (SD = 9.31) of eye tracking data and 15.59% (SD = 8.16) of manually coded data was lost for each participant, $\boldsymbol{t(23.37) = 0.09}$, $\boldsymbol{p = .929}$, $\boldsymbol{\Delta M = 0.00}$, 95% CI $\boldsymbol{[-0.06, 0.07]}$. Additionally, previous research has combined data across these methods to create a single bilingual sample (Byers-Heinlein et al., 2021), further supporting this approach.

Results

Data for each trial were analyzed between 400 and 2000 ms after the onset of the target noun. While standard approaches typically begin analysis at 367 ms after onset of the target noun (Swingley, 2012), we opted to start our analysis window slightly later in order to create consistent 100-ms time bins to use in a growth curve analysis (see below). Trials where the child was inattentive (i.e., looked at the pictures for less than

750 ms during this window) were excluded from the analyses. Children who did not successfully complete at least 2 single-language and 2 code-switched trials were also removed from the analyses. Out of 8 possible trials of each type, children retained for analysis completed an average of 6.87 single-language trials (range = 3 – 8) and 6.63 code-switched trials (range = 4 – 8). To determine if children demonstrated successful comprehension of the target words, we examined the proportion of time that they looked towards the target picture on each trial. This was calculated by dividing the looking time to the target picture by the total time spent looking at either picture. Analyses were conducted using the R statistical language (R Core Team, 2022).

First, we investigated whether children showed comprehension of the noun on each trial type. One-sample, two-sided *t*-tests revealed that children looked significantly above chance ($\mu_0 = 0.5$) to the target picture on both single-language trials, $t(29) = 11.42$, $p < .001$, $M = 0.74$, 95% CI [0.70,0.78], and code-switched trials, $t(29) = 12.03$, $p < .001$, $M = 0.78$, 95% CI [0.73,0.82], indicating a robust ability to understand the target noun in both trial types (see Figure 2).

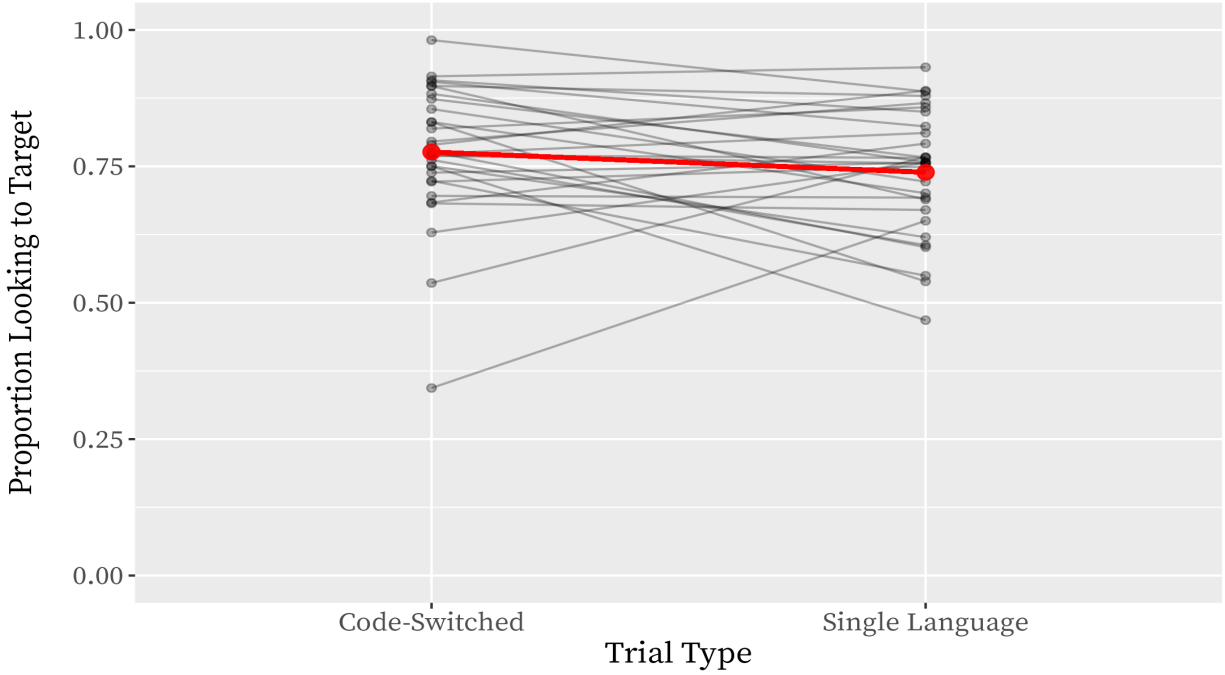


Figure 2. ***Proportion looking to target picture by trial type for all children. The larger red dots and line represent the grand mean. Smaller gray dots and their connecting lines represent the mean values for individual participants.***

We then compared looking time during the two trial types using a paired-samples t-test. The effect of trial type was not statistically significant, $t(29) = 1.49$, $p = .148$,

$M_D = 0.04$, 95% CI $[-0.01, 0.09]$, suggesting that children's comprehension of the noun did not differ between single-language and code-switched trials. Contrary to our prediction that children's comprehension of the target noun would be impaired by the code-switching that preceded it, this result indicated that they were potentially unaffected by the code-switched adjective.

Growth Curve Analysis

The previous analyses, which are typical in this area of research, collapsed infants' data across the entire time window and averaged across trial types to yield two data points per child. However, it has long been recognized in the field that time course data can offer revealing information about children's performance (e.g., Fernald et al., 2001). Analytic techniques such as growth curve analysis (Mirman, 2017) offer an approach to quantify differences in time course, and further allow analysis of trial-level data, thus increasing statistical power. We plotted the time course of our data and then conducted an exploratory growth curve analysis, using the same time window of 400 – 2000ms. Looking-time data were binned in 100ms blocks.

Models were built iteratively. We started with a baseline model with only linear and quadratic time terms and by-participant random effects on both time terms. We then added one additional individual difference variable to the model and compared the two nested models with an analysis of variance. Only variables that significantly improved model fit ($p < .05$) according to a chi-squared test were retained. Intermediary models are available in the supplementary materials. The categorical variables of trial type, testing location, and language dominance were coded using a simple contrast coding scheme. SES and vocabulary size were continuous. We estimated parameter estimate degrees of freedom and *p*-values using Satterthwaite's method.

To address our main research question of the effect of code-switching on children's comprehension, our first exploratory model added trial type to the baseline model described above. We then conducted additional exploratory growth curve models building from this model looking at the potential individual effects of language dominance, testing location, SES, and vocabulary size.

Trial Type

In the growth curve model investigating the effect of trial type, the fixed effects of the final model included trial type, and linear and quadratic time terms. There was a statistically significant main effect of trial type, indicating that, opposite to our prediction, children were more accurate at gazing toward the target picture when hearing code-switched trials compared to single-language trials $t(6,100.82) = -3.43$, $p = .001$, $\hat{\beta} = -0.03$, 95% CI $[-0.05, -0.01]$ (See Table 3 for full results). This result differs from that of the paired-samples *t*-test, which did not find a statistically significant

difference in children's looking between the two trial types.

Table 3. ***Growth curve analysis including trial type.***

	Estimate	95% CI	t	df	p
Fixed effects					
Intercept	0.76	[0.72, 0.79]	43.05	29.42	< .001
Time (Linear)	0.29	[0.14, 0.43]	3.86	29.46	.001
Time (Quadratic)	-0.27	[-0.32, -0.23]	-12.36	29.09	< .001
Trial type	-0.03	[-0.05, -0.01]	-3.43	6,100.82	.001
Random effects		**Variance**			
Participant	Intercept	0.008			
	Time (Linear)	0.154			
	Time (Quadratic)	0.002			

Individual Differences

As previous studies have found some evidence of individual differences in bilingual children's ability to process code-switching (Byers-Heinlein et al., 2021; Potter et al., 2019), we next investigated how such differences may have affected children's performance on this task. Prior to conducting these individual differences analyses, we first quantified the consistency of children's performance, by estimating the reliability of the looking time to each trial type using an intraclass correlation coefficient (ICC), based on a mean-rating, consistent, 2-way random-effects model (Byers-Heinlein et al., 2022). The estimated consistency was 0.19, 95% CI = [–0.24, 0.51] for single-language trials and 0.39, 95% CI = [0.07, 0.64] for code-switched trials. The magnitude of these ICCs was higher than in many other infant studies (Byers-Heinlein et al., 2022), supporting a cautious investigation of individual differences. However, these ICCs could be considered moderate to low on an absolute scale thus reducing statistical power for detecting correlations with other measures of individual differences.

We investigated four individual difference variables: language dominance, testing location (which was also a proxy for language pair), SES, and vocabulary size. We note that the last three variables were interrelated in our dataset: children from Montreal generally came from higher SES backgrounds, $t(12.17) = 2.35$, $p = .037$, $\Delta M = 3.76$, 95% CI [0.27,7.25], and had a larger vocabulary, $t(26.73) = -2.16$, $p = .040$, $\Delta M = -37.76$, 95% CI [−73.56, −1.95], than children from New Jersey. Given our sample size, it was not possible to statistically disentangle these factors. Thus, our approach was to create separate models for each variable to gain some insight into which factor might have the largest explanatory power. We did so by adding each variable to the previous model including trial type as a main effect and in an interaction with trial

type. Here, we focus on the specific effect of these terms. Full results of these models are reported in the supplementary materials. Note that we also used the same approach to explore the potential impact of a fifth individual difference variable – children's exposure to parental code-switching as measured by the Language Mixing Scale. This model did not explain significantly more variance than a base model without the code-switching predictor, and thus we did not interpret this model.

In each of the four models there was a statistically significant main effect of trial type, indicating that, opposite to our prediction, children were more accurate at gazing towards the target picture when hearing code-switched trials compared to single-language trials, whether controlling for language dominance, $t(6{,}101.58) = -3.39$, $p = .001$, $\hat{\beta} = -0.03$, 95% CI $[-0.05, -0.01]$, testing location, $t(6{,}103.15) = -4.67$, $p < .001$, $\hat{\beta} = -0.05$, 95% CI $[-0.07, -0.03]$, SES, $t(6{,}106.01) = -4.75$, $p < .001$, $\hat{\beta} = -0.20$, 95% CI $[-0.28, -0.12]$, or vocabulary, $t(5{,}899.58) = -2.10$, $p = .035$, $\hat{\beta} = -0.05$, 95% CI $[-0.10{,}0.00]$.

We then examined the main effect of each individual difference variable and its interaction with trial type (See Figure 3), and an interesting pattern of results emerged. For language dominance, there was no statistically significant main effect, $t(29.44) = -1.36$, $p = .183$, $\hat{\beta} = -0.05$, 95% CI $[-0.11{,}0.02]$, or interaction with trial type, $t(6{,}101.58) = 0.35$, $p = .727$, $\hat{\beta} = 0.01$, 95% CI $[-0.03{,}0.05]$, suggesting that both children tested in their dominant language and children tested in their non-dominant language performed similarly across trial types. Effects of testing location, SES, and vocabulary showed similar patterns across models. Analyses of testing location revealed that children from Montreal performed similarly on both trial types, whereas children from New Jersey performed better on code-switched than single-language trials $t(6{,}103.14) = -4.16$, $p < .001$, $\hat{\beta} = -0.09$, 95% CI $[-0.13, -0.05]$.

To follow up on the Montreal results, we conducted the pupillometry analyses reported in supplementary materials, which support the main finding that children did not process code-switched and single-language trials differently (these analyses could not be carried out for New Jersey participants, as their data were hand coded from a video recording rather than collected via an eye-tracker). SES analyses showed that children from higher-SES backgrounds performed similarly across trial types whereas children from lower-SES backgrounds performed better on code-switched than single-language trials, $t(6{,}103.72) = 4.04$, $p < .001$, $\hat{\beta} = 0.01$, 95% CI [0.01,0.02]. Finally, children with larger vocabularies performed better across trial types (i.e., looked more to the labeled target in general) than children with smaller vocabularies, $t(28.38) = 2.42$, $p = .022$, $\hat{\beta} = 0.0007$, 95% CI [0.0001,0.0013], but the effect of vocabulary size did not differ significantly as a function of trial type, $t(5{,}896.30) = 0.85$, $p = .396$, $\hat{\beta} = 0.0002$, 95% CI [-0.0002,0.0005].

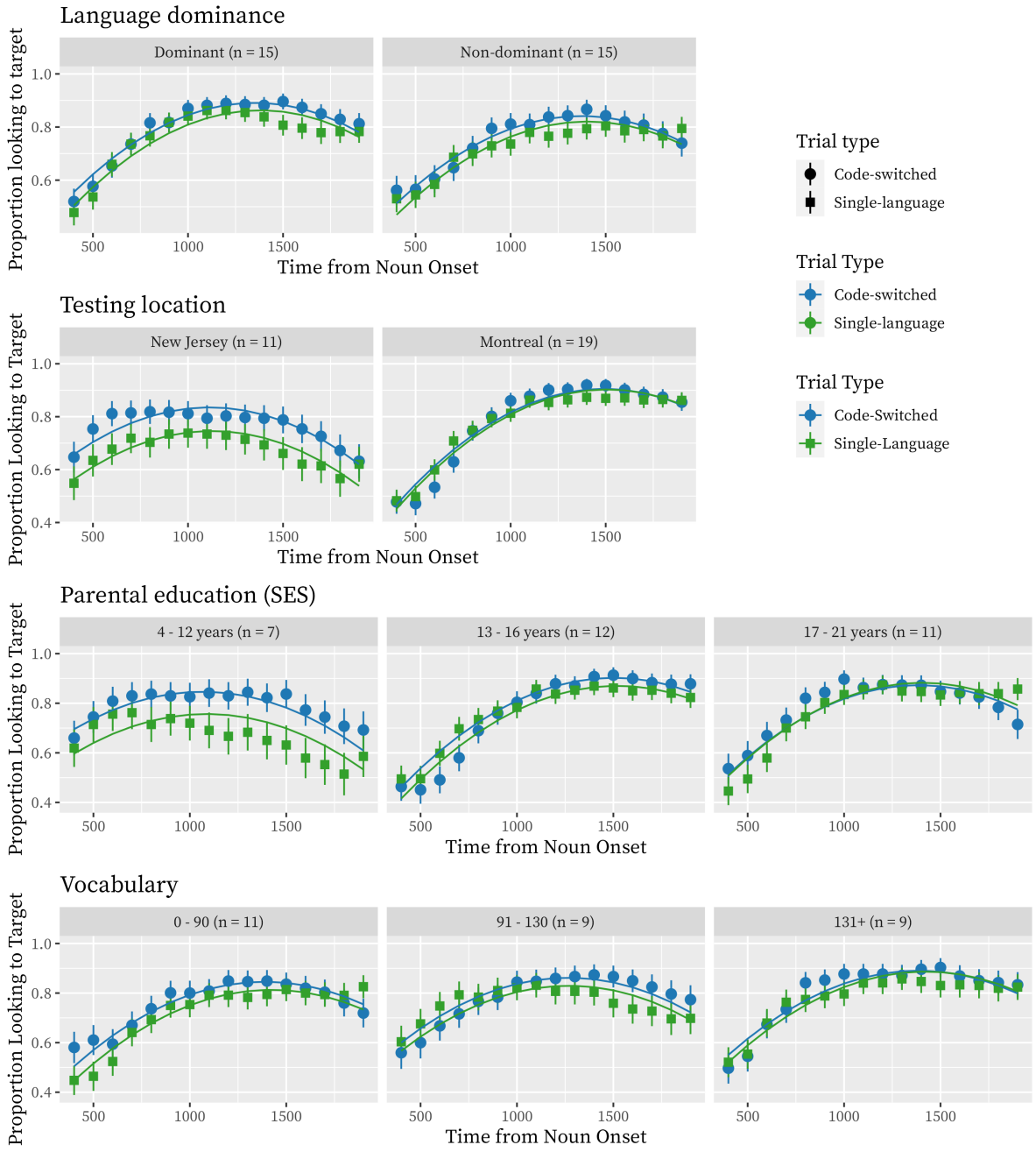


Figure 3. ***Proportion looking to target picture throughout the analysis window. Dots represent means averaged over participants, bars represent ± 1 SEM, and lines represent the growth curve analysis model. SES and vocabulary were included in the model as a continuous variable but have been split into categories for the purposes of visualization. Note that one participant did not have a vocabulary score and was thus excluded from that model.***

These results indicate that individual differences in performance across the two trial types were statistically related to testing location and SES, but not to language dominance or vocabulary size. Spanish-English bilingual children from New Jersey, particularly those whose parents had received a high school education or less (i.e., 12 years or fewer; see Figure 3), performed better on code-switched trials compared to single-language trials, whereas French-English bilingual children and those whose parents had more education performed similarly on the two trial types. Together, the findings show the importance of examining individual differences between participants and samples, as bilingual children's comprehension of these code-switched sentences was not uniform.

Discussion

This study compared bilingual children's comprehension of sentences with code-switching at an uninformative determiner-adjective pair (e.g., "Can you find *le bon* [fr. the good] duck?") to their comprehension of single-language sentences (e.g., "Can you find the good duck?"). We tested 3-year-old bilingual children, including French–English bilinguals in Montreal and Spanish–English bilinguals in New Jersey. We found that bilinguals were, on average, successful at identifying the target noun in both types of sentences, and we did not see evidence that code-switching at an uninformative adjective caused any difficulties in sentence processing. Language dominance did not affect performance, likely because the target noun was always presented in a consistent language, and the switch occurred at the preceding adjective. This finding contrasts with prior reports of dominance effects in studies of children's processing of code-switches (Potter et al., 2019). Surprisingly, we found some evidence that, for certain children, code-switched sentences may have facilitated comprehension relative to single-language sentences. Our experimental design allowed us to evaluate two general accounts of why code-switching impacts speech comprehension. Under the frequency account of code-switch processing, the infrequent nature of code-switching at a determiner-adjective pair should have hindered children's comprehension, perhaps even more so than code-switching at nouns (Byers-Heinlein et al., 2017; Morini & Newman, 2019; Potter et al., 2019). In contrast, under the functional account, children may have been able to seamlessly process code-switching at an uninformative adjective, because they did not need to integrate the meaning of the adjective to identify the target noun. While these two accounts are not mutually exclusive, our results generally support the functional account as children were able to understand the code-switch sentences as well as the single-language sentences. Below, we further discuss why young children's processing was not disrupted by code-switching at uninformative adjectives. Then, we turn to addressing the observed individual differences between participants and communities.

A key aspect of our experimental design was that the determiner-adjective pair in our sentences was uninformative. Children heard sentences with mid-sentence code-switching, as in "Can you find *le bon* [fr. the good] duck?" Critically, the adjective "*bon*" [fr. good] did not add relevant information for identifying the target object, as there was only one duck on the screen. Children typically process the meaning of adjective–noun phrases incrementally (Fernald et al., 2010; Tribushinina & Mak, 2016), but they can "listen through" the adjective to quickly identify the target object when a prenominal adjective is uninformative and does not disambiguate two objects (Thorpe & Fernald, 2006). Following the functional account, code-switching may not be disruptive when the information it carries does not need to be retrieved or integrated into processing. Children may not have experienced a code-switching cost in the current study, because they did not need to process the meaning of the code-switched adjective to identify the target and were therefore able to ignore it.

Similarly, if code-switching is related to prediction processes during language comprehension (e.g., Yacovone et al., 2021), the unexpected code-switch at the adjective might have led to a brief processing slowdown combined with a simultaneous increase in attention (Reuter et al., 2019), effectively canceling each other out in the context of an uninformative adjective. Thus, derailment in children's processing of code-switches may be limited to functionally important words or phrases that require them to integrate the information contained in the switch. Importantly – and in contrast to our study design – adjectives often do carry functional importance in a sentence; for example, the word "heavy" can help distinguish between items of different weights, and "yummy" can refer to a food that is more delicious than another.

To further test this possibility, future studies could compare performance on trials like those in the current study and trials with an informative adjective (e.g., by showing a picture of a big and small duck and examining children's real-time interpretation of the sentence "Do you see *le petit* [fr. the little] duck?"). Under the functional account, sentences with an informative adjective would presumably result in a code-switching cost, because children would no longer be able to "listen through" the code-switched adjective and would potentially need to engage their other language more fully.

While "listening through" could explain why we did not observe a code-switching cost, it does not explain the observed individual differences in children's performance on code-switched and single-language sentences. Our analyses revealed that testing location and SES accounted for significant individual variation in performance across the single-language and code-switched trials, but language dominance and vocabulary size did not. Specifically, children from higher-SES backgrounds performed similarly across trial types; children from lower-SES backgrounds, particularly whose parents had a high school education or less, performed better on code-switched trials than single-language trials, and were all Spanish–English bilinguals in New Jersey.

In our sample, testing location (a proxy for language pair), SES, and vocabulary size were tightly related: French–English children from Montreal had higher vocabularies and were from higher SES backgrounds on average than Spanish–English children from New Jersey. Because of the correlational nature of this finding and the interrelatedness of these variables, it is not possible to pinpoint the factors driving the individual differences we observed, and thus this is a limitation of our study. However, previous studies have reported similar patterns of individual differences in infants from these same communities; one study suggested that Spanish–English children may have slightly weaker skills in real-time language tasks than French–English children (Byers-Heinlein et al., 2021). Following the functional account, if some children were slower to switch between processing their two languages, or if they were less aware of its meaning, it is possible that they were able to "listen through" the uninformative adjective more easily (or under a prediction-based framework, encountered little to no prediction error). However, note that under this explanation, we would have expected vocabulary size to predict performance, which it did not. Rather, SES was a predictor of performance, a variable which has previously been related to children's language development (Fernald et al., 2013; Pace et al., 2017; Pungello et al., 2009). We tentatively suggest that experiential factors related to SES might be driving the observed community differences we observed.

There are also other potentially relevant differences between children that we were not able to directly observe that may have affected infants' performance on our task, which again limits our conclusions. For example, different infants have different experiences with code-switching (Bail et al., 2015; Kremin et al., 2021), which could in turn impact their comprehension of code-switching. The frequency account predicts that bilinguals with frequent exposure to code-switching should experience less disruption in processing compared to bilinguals without frequent exposure to code-switching (Gosselin & Sabourin, 2021; Valdés Kroff et al., 2018). In the context of the current study, experience with code-switching may have been able to build on top of children's ability to "listen through" the uninformative adjective, supporting aspects of both the frequency and the functional account. Indeed, preliminary evidence from a direct observation study suggests that children's experiences hearing code-switching may be somewhat different in the two communities from which we sampled, with Spanish-English bilingual caregivers in New Jersey engaging in more code-switching than English-French bilingual caregivers in Montreal (Kosie et al., 2022). It is also possible that production of code-switching varies by SES within the the two communities we studied, although this has not yet been examined directly. Indeed, we speculate that if Spanish-English bilinguals in New Jersey, particularly those from lower-SES backgrounds, were somewhat more accustomed to hearing code-switching than the French-English bilinguals in Montreal, this could result in the observed "boost" in real-time sentence interpretation – at least in the context of sentences with mid-sentence code-switches at uninformative locations. To address this question, additional

research is needed to directly investigate the relationship between the amount and type of code-switching that bilingual children hear and how they process incoming speech input in two languages.

Finally, this work adds important qualifications to the idea that code-switching engenders processing costs in bilingual children. Our study found that code-switching of uninformative adjectives does not hinder children's comprehension of a subsequent noun and indeed it may have facilitated comprehension, at least for some children. Such facilitatory effects have been reported in the adult literature, whereby a code-switch cued participants that a low-frequency word would be heard, allowing listeners to rapidly identify a labeled target (Tomić & Valdés Kroff, 2021). Parents of young bilinguals code-switch at a variety of syntactic locations and for a variety of reasons, and our results support the idea that certain instances of code-switching do not hinder processing but may even support comprehension and learning (Kremin et al., 2021). It is not scientifically sound to tell parents that code-switching is 'good' or 'bad', and future experiments will need to carefully document young bilinguals' everyday experience with code-switching and evaluate how they process instances of typical and atypical switching.

Conclusion

Code-switching is common in bilingual speech, making it important to understand its effect on children's language comprehension and language learning. Past research has generally found that code-switching leads to processing costs, but in the current study, bilingual children did not show this processing cost. Growth curve analyses revealed that bilinguals showed similar (and in some cases, better) processing of sentences with a code-switch at an uninformative adjective phrase, relative to single-language sentences. These findings demonstrate that linguistic features such as informativeness and location, together with individual-difference variables, may impact how bilingual children process code-switching in natural settings.

References

Abutalebi, J., Brambati, S. M., Annoni, J.-M., Moro, A., Cappa, S. F., & Perani, D. (2007). The Neural Cost of the Auditory Perception of Language Switches: An Event-Related Functional Magnetic Resonance Imaging Study in Bilinguals. *Journal of Neuroscience, 27*(50), 13762–13769. https://doi.org/10.1523/JNEUROSCI.3294-07.2007

Bail, A., Morini, G., & Newman, R. S. (2015). Look at the gato! Code-switching in speech to toddlers*. *Journal of Child Language*, *42*(5), 1073–1101. https://doi.org/10.1017/S0305000914000695

Beatty-Martínez, A. L., Valdés Kroff, J. R., & Dussias, P. E. (2018). From the field to the lab: A converging methods approach to the study of codeswitching. *Languages*, *3*(2), 19. https://doi.org/10.3390/languages3020019

Boudreault, M.-C., Cabirol, É.-A., Trudeau, N., Poulin-Dubois, D., & Sutton, A. (2007). Les inventaires MacArthur du développement de la communication: Validité et données normatives préliminaires. *Revue Canadienne d'orthophonie Et d'audiologie*, *31*(1), 27–37.

Byers-Heinlein, K. (2013). Parental language mixing: Its measurement and the relation of mixed input to young bilingual children's vocabulary size. *Bilingualism; Cambridge*, *16*(1), 32–48. https://doi.org/10.1017/S1366728912000120

Byers-Heinlein, K. (2015). Methods for studying infant bilingualism. In J. W. Schwieter (Ed.), *The Cambridge Handbook of Bilingual Processing* (pp. 133–154). Cambridge University Press. https://doi.org/10.1017/CBO9781107447257.005

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2022). Six solutions for more reliable infant research. *Infant and Child Development*, *31*(n/a), e2296. https://doi.org/10.1002/icd.2296

Byers-Heinlein, K., Jardak, A., Fourakis, E., & Lew-Williams, C. (2021). Effects of language mixing on bilingual children's word learning. *Bilingualism: Language and Cognition*, 1–15. https://doi.org/10.1017/S1366728921000699

Byers-Heinlein, K., Morin-Lessard, E., & Lew-Williams, C. (2017). Bilingual infants control their languages as they listen. *Proceedings of the National Academy of Sciences*, *114*(34), 9032–9037. https://doi.org/10.1073/pnas.1703220114

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test–Fourth Edition*. https://doi.org/10.1037/t15144-000

Dunn, L. M., Dunn, L. M., & Thériault-Whalen, C. M. (1993). *Échelle de vocabulaire en images Peabody: EVIP*. Psycan.

Dunn, L. M., Padilla, E. R., Lugo, D. E., & Dunn, L. M. (1986). *TVIP: Test de vocabulario en imágenes Peabody: Adaptación Hispanoamericana*. American Guidance Service.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., & Bates, E. (2007). *MacArthur-Bates communicative development inventories (CDIs)* (Second). Paul H. Brookes Publishing Company Baltimore, MD.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science, 16*(2), 234–248. https://doi.org/10.1111/desc.12019

Fernald, A., Swingley, D., & Pinto, J. P. (2001). When half a word is enough: Infants can recognize spoken words using partial phonetic information. *Child Development, 72*(4), 1003–1015. https://doi.org/10.1111/1467-8624.00331

Fernald, A., Thorpe, K., & Marchman, V. A. (2010). Blue car, red car: Developing efficiency in online interpretation of adjectivenoun phrases. *Cognitive Psychology, 60*(3), 190–217. https://doi.org/10.1016/j.cogpsych.2009.12.002

Goodz, N. S. (1989). Parental language mixing in bilingual families. *Infant Mental Health Journal, 10*(1), 25–44. https://doi.org/10.1002/1097-0355(198921)10:1<25::AID-IMHJ2280100104>3.0.CO;2-R

Gosselin, L., & Sabourin, L. (2021). Lexical-semantic processing costs are not inherent to intra-sentential code-switching: The role of switching habits. *Neuropsychologia*, 107922. https://doi.org/10.1016/j.neuropsychologia.2021.107922

Guzzardo Tamargo, R. E., Valdés Kroff, J. R., & Dussias, P. E. (2016). Examining the relationship between comprehension and production processes in code-switched language. *Journal of Memory and Language, 89*, 138–161. https://doi.org/10.1016/j.jml.2015.12.002

Heredia, R. R., & Altarriba, J. (2001). Bilingual Language Mixing: Why Do Bilinguals Code-Switch? *Current Directions in Psychological Science*, *10*(5), 164–168. https://doi.org/10.1111/1467-8721.00140

Jackson-Maldonado, D., Thal, D. J., & Fenson, L. (2003). *MacArthur Inventarios del Desarrollo de Habilidades Comunicativas: User's guide and technical manual*. Brookes Pub.

Kaushanskaya, M., & Crespo, K. (2019). Does Exposure to Code-Switching Influence Language Performance in Bilingual Children? *Child Development*, *90*(3), 708–718. https://doi.org/10.1111/cdev.13235

Kosie, J. E., Fibla, L., Tsui, R. K.-Y., Martinez, T., Byers-Heinlein, K., & Lew-Williams, C. (2022). *Infants' exposure to language switching in bilingual homes across two communities*.

Kremin, L. V., Alves, J., Orena, A. J., Polka, L., & Byers-Heinlein, K. (2021). Code-switching in parents' everyday speech to bilingual infants. *Journal of Child Language*, 1–27. https://doi.org/10.1017/S0305000921000118

Libertus, M. E., Odic, D., Feigenson, L., & Halberda, J. (2015). A Developmental Vocabulary Assessment for Parents (DVAP): Validating Parental Report of Vocabulary Size in 2- to 7-Year-Old Children. *Journal of Cognition and Development, 16*(3), 442–454. https://doi.org/10.1080/15248372.2013.835312

MacSwan, J. (2012). Code-Switching and Grammatical Theory. *The Handbook of Bilingualism and Multilingualism*, 323.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals. *Journal of Speech Language and Hearing Research, 50*(4), 940. https://doi.org/10.1044/1092-4388(2007/067)

Marino, C., Bernard, C., & Gervain, J. (2020). Word Frequency Is a Cue to Lexical Category for 8-Month-Old Infants. *Current Biology, 30*(8), 1380–1386.e3. https://doi.org/10.1016/j.cub.2020.01.070

Mirman, D. (2017). *Growth curve analysis and visualization using R*. Chapman and Hall/CRC.

Morini, G., & Newman, R. S. (2019). Dónde está la ball? Examining the effect of code switching on bilingual children's word recognition. *Journal of Child Language*, 1–11. https://doi.org/10.1017/S0305000919000400

Ng, S., Gonzalez, C., & Wicha, N. Y. Y. (2014). The fox and the cabra: An ERP analysis of reading code switched nouns and verbs in bilingual short stories. *Brain Research, 1557*, 127–140. https://doi.org/10.1016/j.brainres.2014.02.009

Nilep, C. (2006). "Code Switching" in Sociocultural Linguistics. *Colorado Research in Linguistics, 19*(1). https://doi.org/10.25810/hnq4-jv62

Pace, A., Luo, R., Hirsh-Pasek, K., & Golinkoff, R. M. (2017). Identifying pathways between socioeconomic status and language development. *Annual Review of Linguistics, 3*, 285–308. https://doi.org/10.1146/annurev-linguistics-011516-034226

Potter, C. E., Fourakis, E., Morin-Lessard, E., Byers-Heinlein, K., & Lew-Williams, C. (2019). Bilingual toddlers' comprehension of mixed sentences is asymmetrical across their two languages. *Developmental Science, 22*(4), e12794. https://doi.org/10.1111/desc.12794

Pungello, E. P., Iruka, I. U., Dotterer, A. M., Mills-Koonce, R., & Reznick, J. S. (2009). The effects of socioeconomic status, race, and parenting on language development in early childhood. *Developmental Psychology, 45*(2), 544. https://doi.org/10.1037/a0013917

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reuter, T., Borovsky, A., & Lew-Williams, C. (2019). Predict and redirect: Prediction errors support children's word learning. *Developmental Psychology, 55*(8), 1656–1665. https://doi.org/10.1037/dev0000754

Salig, L. K., Valdés Kroff, J. R., Slevc, L. R., & Novick, J. M. (2023). Linking frequency to bilingual switch costs during real-time sentence comprehension. *Bilingualism: Language and Cognition,* 1–16. https://doi.org/10.1017/S1366728923000366

Swingley, D. (2012). The Looking-While-Listening Procedure. In E. Hoff (Ed.), *Research Methods in Child Language* (pp. 29–42). Wiley-Blackwell. https://doi.org/10.1002/9781444344035.ch3

Thorpe, K., & Fernald, A. (2006). Knowing what a novel word is not: Two-year-olds "listen through" ambiguous adjectives in fluent speech. *Cognition, 100*(3), 389–433. https://doi.org/10.1016/j.cognition.2005.04.009

Tomić, A., & Valdés Kroff, J. R. (2021). Expecting the unexpected: Code-switching as a facilitatory cue in online sentence processing. *Bilingualism: Language and Cognition,* 1–12. https://doi.org/10.1017/S1366728921000237

Tribushinina, E., & Mak, W. M. (2016). Three-year-olds can predict a noun based on an attributive adjective: Evidence from eye-tracking. *Journal of Child Language; Cambridge, 43*(2), 425–441. https://doi.org/http://dx.doi.org.libezproxy.concordia.ca/10.1017/S0305000915000173

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition, 73*(2), 89–134. https://doi.org/10.1016/S0010-0277(99)00032-3

Valdés Kroff, J. R., Guzzardo Tamargo, R. E., & Dussias, P. E. (2018). Experimental contributions of eye-tracking to the understanding of comprehension processes while hearing and reading code-switches. *Linguistic Approaches to Bilingualism, 8*(1), 98–133. https://doi.org/10.1075/lab.16011.val

van Hell, J. G., Fernandez, C. B., Kootstra, G. J., Litcofsky, K. A., & Ting, C. Y. (2018). Electrophysiological and experimental-behavioral approaches to the study of intra-sentential code-switching. *Linguistic Approaches to Bilingualism, 8*(1), 134–161. https://doi.org/10.1075/lab.16010.van

Vaughan-Evans, A., Parafita Couto, M. C., Boutonnet, B., Hoshino, N., Webb-Davies, P., Deuchar, M., & Thierry, G. (2020). Switchmate! An Electrophysiological Attempt to Adjudicate Between Competing Accounts of Adjective-Noun Code-Switching. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.549762

Venker, C. E., Pomper, R., Mahr, T., Edwards, J., Saffran, J., & Weismer, S. E. (2020). Comparing Automatic Eye Tracking and Manual Gaze Coding Methods in Young Children with Autism Spectrum Disorder. *Autism Research, 13*(2), 271–283. https://doi.org/10.1002/aur.2225

Yacovone, A., Moya, E., & Snedeker, J. (2021). Unexpected words or unexpected languages? Two ERP effects of code-switching in naturalistic discourse. *Cognition, 215*, 104814. https://doi.org/10.1016/j.cognition.2021.104814

Zeller, J. P. (2020). Code-Switching Does Not Equal Code-Switching. An Event-Related Potentials Study on Switching From L2 German to L1 Russian at Prepositions and Nouns. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.01387

Data, code, and materials availability statement

All stimuli, data, and analysis scripts for the current study are available via the Open Science Framework at https://doi.org/10.17605/OSF.IO/ECQWR.

Ethics statement

Ethics approval was obtained from the Concordia University Human Research Ethics Committee ("Monolingual and Bilingual Language Development"; approval #10000493) and the Princeton University Institutional Review Board ("Language learning and Communication"; approval #7117), and parents provided informed consent prior to their child's participation.

Authorship and Contributorship Statement

The authors made the following contributions. Lena V. Kremin: Formal analysis, Writing – Original draft, Visualization; Amel Jardak: Conceptualization, Methodology, Formal analysis, Investigation, Writing – Original Draft; Casey Lew-Williams: Conceptualization, Methodology, Writing – Review & Editing, Supervision, Project

administration, Funding acquisition; Krista Byers-Heinlein: Conceptualization, Methodology, Writing – Review & Editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

This work was supported by a grant to KBH from the Natural Sciences and Engineering Research Council of Canada (402470-2011), grants to CLW and KBH from the U.S. National Institute of Child Health and Human Development (R01HD095912, R03HD079779), a grant to CLW from the Speech-Language-Hearing Foundation, support for KBH from the Concordia University Research Chairs program, and fellowships to LVK and AJ from the Fonds de Recherche du Québec–Société et Culture. We thank the many members of the Concordia Infant Research Lab and the Princeton Baby Lab who assisted with testing participants, and the children and parents who participated.

License

Word learning in 14-month-old monolinguals and bilinguals: Challenges and methodological opportunities

Ana Maria Gonzalez-Barrero[1]
Rodrigo Dal Ben[2]
Hilary Killam
Krista Byers-Heinlein

Department of Psychology, Concordia University, Canada
Centre for Research on Brain, Language and Music, Canada

Abstract: Infants can learn words in their daily interactions early in life, and many studies have demonstrated that they can also learn words from brief in-lab exposures. While most studies have included monolingual infants, less is known about bilingual infants' word learning and the role that language familiarity plays in this ability. In this study we examined word learning in a large sample (up to $N = 148$) of bilingual and monolingual 14-month-olds using a preferential looking paradigm. Two novel words were presented within sentence frames in one language (single-language condition) or two languages (dual-language condition). We predicted that infants would learn both words, and would exhibit better learning when they were more familiar with the sentence frame language. Using a traditional analytic approach (*t*-tests) and a standard linear regression, we found weak evidence that children learned one of the two words. However, contrary to our prediction, in a minority of conditions infants may have learned better when stimuli were presented in sentence frames in a less familiar language. We also conducted updated analyses using mixed-effects linear regression models, which did not support the conclusion that infants learned any of the words they encountered, regardless of the familiarity of the sentence frame language. We discuss these results in relation to prior work and suggest how open science practices can contribute to more reliable findings about early word learning.

Keywords: word learning; infants; bilingualism; open science.

Corresponding author: Ana Maria Gonzalez-Barrero, Department of Psychology, Concordia University, 7141 Sherbrooke St. West, Montreal, QC, H4B 1R6, Canada. E-mail: ana.gonzalez@mail.mcgill.ca

ORCID ID(s): Ana Maria Gonzalez-Barrero (https://orcid.org/0000-0002-2120-6329), Rodrigo Dal Ben (https://orcid.org/0000-0003-2185-8762), Hilary Killam (https://orcid.org/0000-0002-3080-0610), Krista Byers-Heinlein (https://orcid.org/0000-0002-7040-2510)

Citation: Gonzalez-Barrero, A. M., Dal Ben, R., Killam, H., & Byers-Heinlein, K. (2023). Word learning in 14-month-old monolinguals and bilinguals: Challenges and methodological opportunities. *Language Development Research, 3*(1), 277—317. http://doi.org/10.34842/3vw8-k253

---

[1] Ana Maria Gonzalez-Barrero is now at School of Communication Sciences and Disorders, Faculty of Health, Dalhousie University, Nova Scotia, Canada.
[2] Rodrigo Dal Ben is now at Ambrose University, Psychology Program, Alberta, Canada.

Introduction

Word learning is a complex process that begins to unfold over the first two years of life. Past research, using a variety of experimental designs, has provided important insights into word learning in both monolingual and bilingual infants (e.g., Fennell & Byers-Heinlein, 2014; Graf Estes, 2014; Kalashnikova et al., 2018; Mattock et al., 2010; Schafer & Plunkett, 1998; Singh et al., 2018; Taxitari et al., 2020; Werker et al., 1998; Woodward et al., 1994). However, a new understanding of research best practices highlights the limitations of our traditional methodological and statistical approaches (Bergmann et al., 2018; Oakes, 2017). In this manuscript, we present a case study using both traditional and more sensitive analytic techniques to examine word learning in the lab with a large sample of 14-month-old monolinguals and bilinguals ($N = 148$). Following prior research (Fennell & Byers-Heinlein, 2014; Fennell & Waxman, 2010), we presented infants with novel words embedded into sentence frames, and then tested their learning in a looking time paradigm. We discuss whether infants of this age were able to learn new words in this context, and how our traditional research practices in early language acquisition can be improved to produce more reliable and reproducible findings.

Word Learning

Researchers have used experimental tasks to study word learning for more than 40 years (e.g., Carey & Bartlett, 1978). Dozens of studies have shown that infants and children can learn new words in the lab (Dal Ben et al., 2019; Gonzalez-Gomez et al., 2013; Kalashnikova et al., 2015; Ramon-Casas et al., 2009; Shukla et al., 2011; Singh et al., 2018; Tsui et al., 2019; Tsuji et al., 2020b; Yu & Smith, 2011). They have also shown that infants' ability to make initial mappings between words and their referents, as well as their ability to retain these mappings, are affected by multiple factors related to the nature of the word learning task, including the particular stimuli used and how the words are encountered (Burnham et al., 2018; Hirsh-Pasek et al., 2000; Kucker et al., 2015; McMurray et al., 2012; Werker & Curtin, 2005), as well as other factors like language background (Tsui et al., 2019) and vocabulary size (Werker et al., 2002). Word learning skills also seem to improve with age (Frank et al., 2021). As early as 6 months old, everyday language experience supports infants' ability to associate labels with referents such as food and body parts (Bergelson & Swingley, 2012), and at this age there is also evidence that infants can learn new words in the lab (Shukla et al., 2011). Some have argued, however, that there is a qualitative change in infants' word comprehension abilities that occurs just after their first birthday (Bergelson, 2020). Indeed, there are many more reports of infants aged 12 months and older showing successful word learning in the lab than reports of younger children (e.g., Graf Estes et al., 2007; Lany,

2014; Yin & Csibra, 2015). Our study tested infants at an age where basic laboratory word learning is thought to be relatively robust: 14 months.

Successful word learning in the lab has been reported in different contexts. For example, Woodward et al. (1994) found that 13-month-old monolingual infants mapped a novel word to its referent after only nine encounters with the word–referent pair presented by a live experimenter. More recently, Chen and colleagues (2020) used a similar paradigm and found that monolingual 20-month-olds could learn a native and a foreign word after only 6 encounters with the word–referent pair. Lab studies using more stripped-down tasks (i.e., without live social interaction) have also shown successful word learning. For instance, 15-month-olds learned two novel words in a preferential looking paradigm where isolated novel words (e.g., "*bard*") and pictures of novel objects were paired, with no social agents or social support (Schafer & Plunkett, 1998). In addition, using the Switch task, wherein infants are habituated to two novel word–object pairings and then presented with a mismatch at test, similar results were found with 14-month-old monolinguals (Werker et al., 1998) and bilinguals (Byers-Heinlein et al., 2013).

For slightly older monolingual (15-month-olds, Fennell & Waxman, 2010) and bilingual (17-month-olds, Fennell & Byers-Heinlein, 2014) infants, sentence frames have been shown to enhance word learning and recognition. For instance, monolingual 18-month-olds showed faster recognition of familiar words presented in sentence frames than in isolation (Fernald & Hurtado, 2006). Moreover, while 14-month-olds often find minimal pair word learning challenging (Stager & Werker, 1997), monolingual 14-month-olds successfully mapped a minimal pair (*bin* and *din*) to objects during the Switch task when the words were embedded in sentence frames (Fennell & Waxman, 2010). Similarly, 16-month-old bilinguals, whose languages shared linguistic similarities (e.g., French and Spanish), mapped a minimal pair (*tola* and *dola)* to objects when words were embedded in sentence frames and presented in a live interaction experiment (Havy et al., 2016). Seventeen-month-old monolinguals and French–English bilinguals also learned minimal pair labels (*kem* and *gem*) embedded into sentences that were produced by a speaker that matched their language background (Fennell & Byers-Heinlein, 2014). Sentence frames may support word learning by providing familiar linguistic context, highlighting the referential nature of the word learning task, and decreasing infants' cognitive load. This information might be particularly useful for bilingual infants, who may use sentential frames to navigate between languages (Fennell & Byers-Heinlein, 2014; Fernald & Hurtado, 2006; Havy et al., 2016).

Current Study

The current study extended previous research by investigating word learning just after infants' first birthdays. We asked if 14-month-old infants would successfully learn new words embedded in sentence frames. Moreover, we were interested in the role of infants' language background, specifically whether they were growing up in a monolingual or a bilingual environment. Despite deploying similar mechanisms for word learning (Byers-Heinlein et al., 2013; Byers-Heinlein & Werker, 2009; Kandhadai et al., 2017), infants growing up bilingual are exposed to unique input that may impact their language development (Fennell et al., 2007; Graf Estes & Hay, 2015). For instance, bilingual infants often hear interlocutors alternate between two languages in the same contexts (Place & Hoff, 2011), especially when bilingual parents teach new words (Byers-Heinlein, 2013; Kremin et al., 2022). Recent studies indicate that some types of language alternation make word learning challenging (Byers-Heinlein et al., 2022), although other evidence suggests that many instances of parental code-switching are supportive for learning (Kremin et al., 2022).

We presented monolingual and bilingual infants with novel words embedded in sentence frames that differed in linguistic familiarity. Specifically, we presented 14-month-olds pictures of novel objects paired with the dissimilar-sounding novel words "*kem*" and "*bos*" embedded in English and/or French sentence frames. Our study had two training conditions. In the single-language condition, both words were presented in the same language (either in English or in French sentence frames), and in the dual-language condition each word was presented in a different language (one in English sentence frames and one in French sentence frames). After training, infants were tested in a preferential looking paradigm, where they saw both novel objects side-by-side and heard one of the words in isolation. Infants came from one of three backgrounds: (a) monolingual English *or* French, (b) bilingual English *and* French, or (c) bilingual English *or* French *and* another language. That is, all infants had exposure to one or both of the sentence frame languages (English and French), but to varying degrees, as bilingual infants are rarely perfectly balanced in their exposure to each of their languages (i.e., they typically have a dominant and a non-dominant language; Byers-Heinlein et al., 2019). By including infants from these diverse language backgrounds, we could examine the effects of bilingualism as well as infants' familiarity with the sentence frame languages.

Given the bulk of evidence from the published literature that 14-month-old monolinguals and bilinguals can successfully learn new words in the lab, we expected that at least under some conditions, infants would also be successful in our task. More specifically, we expected that the more familiar infants were with the language of the sentence frame, the better they would learn the novel words. For instance, we expected that word learning would be easier for bilingual infants when they heard the sentence frame in

their dominant rather than in their non-dominant language, and infants should have the most difficulty learning a new word embedded in foreign language sentence frames (e.g., an English monolingual infant hearing a French sentence).

Building on previous research showing that vocabulary size (Werker et al., 2002) and attention (Kannass & Oakes, 2008; Yu & Smith, 2012) can influence infants' word learning, we also investigated these additional variables to provide a more complete account of our findings. Thus, we also explored whether infants would show better learning as a function of how many words they knew in the sentence frame language (e.g., how many words they knew in English), and whether attention during the training phase contributed to successful word learning.

This project began in 2012, and the combination of different language backgrounds and conditions was originally conceptualized as forming a set of 7 different experiments (see Table 1).[3] Following past studies, we had planned a sample size of 16 infants per condition (see Oakes, 2017, for evidence that this sample size is typical of many infant experiments, although a recent meta-analysis has revealed that this often yields underpowered experiments; Bergmann et al., 2018). However, after 7 years of data collection (2012–2019), and despite collecting data from 288 infants (many of whom ultimately had to be excluded from analyses, discussed further below), we were able to achieve our target sample for only some of the experiments, and thus chose to terminate data collection. We note that by this point the last author of this paper (the Principal Investigator) was the only researcher still in the lab from the time the experiment began. A subset of these data from monolingual infants—who were substantially easier to recruit—was published by da Estrela and Byers-Heinlein (2016), who designed the experimental approach and created the stimuli. They reported an experiment under which monolinguals learned the novel words, as well as two experiments in which they

[3] Our original intention was to investigate word learning in French–English bilinguals, building from a series of word learning studies from that time (Fennell & Byers-Heinlein, 2014; Fennell et al., 2007; Mattock et al., 2010). However, recruitment of French–English bilinguals was slow, and we were turning away many interested families with other language backgrounds. We thus expanded our research design to collect data from monolinguals as well as bilingual infants learning French/English and an additional language. The categorization of infants as monolingual or bilingual (rather than taking a continuous approach to language exposure) was consistent with the literature at the time. We prioritized testing infants in the dual language condition (the first condition we designed), and additionally tested monolinguals and French–English bilinguals in the single-language condition. As there are many more monolinguals in our community than any of the groups of bilinguals, these infants were tested in the greatest number of conditions. It was expected that these infants' different relationships with the sentence frame languages might provide some additional insight into factors – such as familiarity – that could influence early word learning, while allowing us to increase the number of infants tested and accommodate a wider range of families.

failed to learn. In retrospect, we note that all of the studies reported back then were underpowered, which could lead to spurious findings (Oakes, 2017). We expand on this point in the Discussion.

Table 1. ***Examples of Infants' Familiarity with Sentence Frame Languages***

Language Group	Experiment Number	Language Background	Infants' Most Familiar Sentence Frame Language (e.g., English; *Look! It's the Bos*!)	Infants' Least Familiar Sentence Frame Language (e.g., French; *Regarde! C'est le Kem!)*
		Dual-Language Condition		
Bilinguals	1	L1 English L2 French	Dominant	Non-Dominant
	2	L1 English L2 Other	Dominant	Foreign
	3	L1 Other L2 English	Non-Dominant	Foreign
Monolinguals[a]	4	L1 English	Dominant/Native	Foreign
		Single-Language Condition		
Bilinguals	5	L1 English L2 French	Dominant	NA
Monolinguals	6	L1 English	Dominant/Native	NA
Monolinguals[a]	7	L1 French	NA	Foreign

Note. In these examples English is the most familiar language and French is the least familiar language. The relationships are reversed when French is the most familiar language. L1 = Infants' dominant (or native in the case of monolinguals) language; L2 = Infants' non-dominant language.

[a]Monolingual infants included in a prior study.

The experiments presented here were conceptualized before new approaches (e.g., large-scale collaborations; ManyBabies Consortium, 2020) and articles calling for better research practices were widely disseminated in the field of developmental psychology (Bergmann et al., 2018; Bishop, 2020; Oakes, 2017; Schott et al., 2019), although such ideas were being discussed in some other fields (Button et al., 2013; Ioannidis, 2005; John et al., 2012; Simmons et al., 2011). However, for different reasons, most notably the slow pace of infant data collection (especially with bilingual infants), we found ourselves analyzing our data after improved research practices were becoming more common in infant research and the field of bilingualism. This laid bare some problematic characteristics of our original approach, which would likely have characterized many published studies in the field: it was not pre-registered; it had small sample sizes per experimental group; planned statistical analyses focused on small individual experiments rather than the dataset as a whole; there was potential for undisclosed flexibility in the analysis; and monolingualism and bilingualism were defined categorically in a way that ultimately excluded many participants who were tested (see Byers-Heinlein, 2015; Kremin & Byers-Heinlein, 2021; Luk, 2015; Luk & Bialystok, 2013, for a longer discussion of categorical versus continuous approaches to bilingualism).

Our conundrum raises an important question for studies with a long gap between study planning and data analysis: when should researchers stick with their original plan that is consistent with the rest of the literature, and when should they use updated approaches such as combined analyses that yield larger sample sizes, advanced statistical methods, and open science practices? We have ultimately decided to take both paths at once, in order to better understand how we should conceptualize older versus newer research practices in the context of the literature on infant word learning. In what follows, we first present our planned analysis (which we refer to as the traditional approach) and then a re-analysis of our data using a more sensitive technique (which we refer to as the updated approach). Finally, we discuss how the use of traditional versus updated approaches can affect our conclusions about infant experimental word learning tasks, contributing to the discussion on how to improve practices in infant research.

Method

Analytic Approaches

We report two analytic approaches. In the traditional approach, we used one-sample, two-tailed *t*-tests against chance to test word learning for each experiment in each condition, following da Estrela and Byers-Heinlein (2016). In the updated approach, we analyzed all experiments in aggregate (the full sample), using mixed-effects models. Critically, both analytic approaches used the same window of analysis, which began

200ms after the onset of the first iteration of the target word and lasted until the end of the testing trial, 10000ms (see Design section). The 200-ms shift was to account for the time it takes infants to initiate an eye movement (Canfield et al., 1997). The total length of the analysis window was 6800ms. Both approaches were implemented in R (R Core Team, 2020) and all data and scripts are available at https://osf.io/upy7f.

Participants

A total of 288 infants were tested between August 2012 and July 2019. This study was conducted in Montreal, Canada, a multicultural city where a high proportion of children are raised in a bilingual environment (Schott et al., 2022). Following exclusion criteria from prior studies (Byers-Heinlein et al., 2021; Mattock et al., 2010; Tsuji et al., 2020b), we excluded infants born premature (i.e., < 37 weeks of gestation, $n = 10$), with low birth weight (i.e., < 2500 grams, $n = 11$), with major health issues ($n = 1$), and those who were too fussy or inattentive to complete the study (for example, children who cried extensively during the experiment were considered fussy and children who refused to look at the screen were considered inattentive; $n = 44$).

We also excluded infants due to technical problems (e.g., connection problems with the eye-tracker; $n = 17$), experimenter error ($n = 4$), parental interference during the experimental portion of the study ($n = 2$), and those without enough looking data obtained from testing trials ($n = 7$). We defined enough looking data as at least 750ms of looking time during the specified windows of analyses for testing trials (following da Estrela & Byers-Heinlein, 2016), to ensure at least minimal attention was paid to the task, thus we excluded trials with less than 750ms of total looking from our analyses.

In addition, we excluded bilingual infants who were not exposed to both languages from birth or for whom age of acquisition was not reported ($n = 29$), bilingual infants who did not meet the study's language criteria, only discovered once infants participated in the study and parents completed the detailed language exposure questionnaire (i.e., exposure to a second language did not reach at least 25%, $n = 26$; see Rocha-Hidalgo & Barr, 2022, for a discussion of bilingualism criteria used in infant studies), infants who were not exposed to the target languages ($n = 3$), or children who were regularly exposed to 3 languages ($n = 11$). Infants who did not have at least one testing trial with adequate looking data per target word (i.e., at least one valid testing trial for "*kem*" and one for "*bos*", $n = 13$) were excluded from the traditional approach. We return to the issue of this reduction of sample size due to exclusions in the Discussion.

The final sample for the traditional approach included 110 14-month-olds (age range: 13 months and 16 days – 15 months and 12 days, Mean: 14 months and 12 days, *SD*: 13.6

days; 57 females) from diverse language backgrounds. Monolingual infants (n = 50) were exposed to one language, either English or French, 90% of the time or more. Bilingual infants were exposed at least 25% of the time to each of two languages, and less than 20% to a third language. We included bilingual infants exposed to English *and* French (n = 35) and bilingual infants exposed to English *or* French and another language (n = 24). Bilingual infants in all studies were exposed to at least one of the sentence frame languages (English and French) since birth. Participants' demographic characteristics are presented in Table 2.

The final sample for the updated approach consisted of 148 infants[4], a 35% increase in sample size compared to the traditional approach. This included all infants from the traditional approach (n = 110). It also included infants who had been excluded from the traditional approach because their language exposure fell outside the criteria established for bilingualism or monolingualism, except for one infant who did not have at least 750ms of looking time during testing trials (n = 25). We also included the 13 infants who had been excluded from the traditional approach for not having at least one valid test trial per novel word. These additional infants could be included in the updated approach because we treated language exposure continuously rather than categorically, and because mixed effects models are able to handle missing data.

Stimuli

We used the same stimuli and general procedure as da Estrela and Byers-Heinlein (2016). All our stimuli are openly available at https://osf.io/g6nrv. The visual stimuli had been used in prior research examining word learning in monolingual and bilingual infants (Byers-Heinlein et al., 2013; Curtin et al., 2009; Fennell et al., 2007; Fennell & Waxman, 2010; Werker et al. 1998; Werker et al., 2002). The auditory stimuli were recorded in our lab, and were originally chosen from the stimuli used in other previous studies of minimal pair word learning in French–English bilinguals: *bos* had been used by Mattock et al. (2010) and *kem* had been used by Fennell and Byers-Heinlein (2014). The two words do not overlap in sound and contain phonemes that are produced similarly across Canadian French and Canadian English (for a more complete comparison of the

[4] Our final samples for both the traditional and updated approaches included 28 monolingual infants whose data were previously published as a subsample of this larger dataset (Experiment 1 and Experiment 2 from da Estrela and Byers-Heinlein, 2016), which we reanalyzed in Studies 4 and 7 (see Table 1). Total sample sizes from both the traditional (n = 110) and updated approach (N = 148) provide our study with more than 80% statistical power to detect moderate to low effect sizes like the one estimated by a meta-analysis of studies with 12- to 16-months-old infants learning words in the Switch Task (d = 0.33; Tsui et al., 2019). However, individually, experiments 1–7 (n = 10–19) were underpowered. The full power analysis is available at: https://osf.io/upy7f.

realization of the relevant speech sounds in each language, see Fennell & Byers-Heinlein, 2014; Mattock et al., 2010).

Table 2. ***Infant Demographic Characteristics by Language Group***

Language Group	*n*	Mean Age in Months *(SD)*	Age Range	Sex	*n* per Language Dominance	Mean % Language Exposure
Traditional Approach						
Mono-lingual	50	14m 13d *(13.3d)*	13m 17d–15m 7d	48% female	24 English native 26 French native	98% EN 98.5% FR
Bilingual English–French	36	14m 14d *(14.9d)*	13m 16d–15m 12d	56% female	20 English dominant 16 French dominant	65% EN & 33% FR 61% FR & 37% EN
Bilingual English or French & Other Language	24	14m 6d *(11.1d)*	13m 24d–14m 29d	58% female	5 English dominant 5 French dominant 14 dominant in another language (6 with English and 8 with French as non-dominant language)	64% EN & 36% OT 61% FR & 39% OT 64% OT & 36% EN or FR
Additional Infants – Updated Approach						
38 infants included in the updated approach		14m 10d *(12.9d)*	13m 22d–15m 12d	56% female	24 English dominant 14 French dominant 3 dominant in another language (1 with English and 2 with French as non-dominant language)	78% EN, 20% FR, & 2% OT 75% FR, 19% EN, & 6% OT 67% OT, 33% EN and/or FR

Note. *M* = mean; *SD* = Standard Deviation, m = months, d = days, EN = English, FR = French, OT = Other Language. The percentage of language exposure does not add to 100% in some cases, since some infants in monolingual or bilingual groups had a small amount of exposure to other languages.

Visual stimuli consisted of two novel objects: a crown shape and a molecule shape (Figure 1). Target words (*bos* and *kem*) were presented embedded in English and/or French sentence frames (training) or in isolation (test). Across experimental conditions the molecule shape was always labelled with the novel word *kem* and the crown shape was always labelled with the novel word *bos*. Three unique tokens/recordings of each target word were used during training, always favouring the natural flow of the auditory stimuli. Identical tokens for the target words were used across all conditions, on both English and French sentences, which was accomplished through cross-splicing tokens that were pronounced in a way that was neither distinctly English nor distinctly French (according to an informal survey of speakers of each language). There were 3 sentence frames used in English ("Look, it's the ___!"... "Do you see the ___?"... "I like the ___!") and 3 in French ("Regarde, c'est le ___!"... Vois-tu le ___?"... "J'aime le ___!"). The novel words (i.e., *bos* and *kem*) were always presented in a sentence-final position to increase their salience (Echols & Newport, 1992; Fernald & Hurtado, 2006), and to support infants in segmenting out the target word even when the sentence frame was less familiar (Seidl & Johnson, 2006). Sentences were matched on length and prosody to minimize differences across the stimuli and were selected to ensure that the stimuli sounded as natural as possible. There were no sentence frames used in the test phase, and so the exact same recordings were used for French and English. The tokens used for the test phase were different from the ones used in the training phase. All stimuli were recorded by a native bilingual English–French female using infant-directed speech.

Auditory and visual stimuli were combined into videos to create training and test trials. Training trials presented the target object looming against a black background. The visual stimulus appeared in silence for the first 1.5 seconds, followed by 8 seconds where it was accompanied by an auditory stimulus with the target novel word embedded in either a French or English sentence, followed finally by 1.5 seconds of silence. Three sentences were presented during each training trial (e.g., "Look, it's the *kem*!"... "Do you see the *kem*?"... "I like the *kem*!"), with an interval of 1.5 seconds of silence between them. The duration of each training trial was approximately 11 seconds. During Test trials, visual stimuli (i.e., the crown-shaped and molecule-shaped objects) were presented side by side for the entire duration of the trial (≈ 10 seconds). During the first 3 seconds, visual stimuli were presented in silence, then isolated target words were played three times (e.g., "*Kem*!"..."*Kem*!"... "*Kem*!") with 1.5 seconds of pause between repetitions. Visual stimuli remained on the screen for a final 1.5 seconds of silence, before a new test trial began. Test trials were presented in one of four counterbalanced orders, which were identical across conditions. Stimuli are available at https://osf.io/g6nrv. Figure 1 shows the stimuli and timeline during an example training and an example test trial.

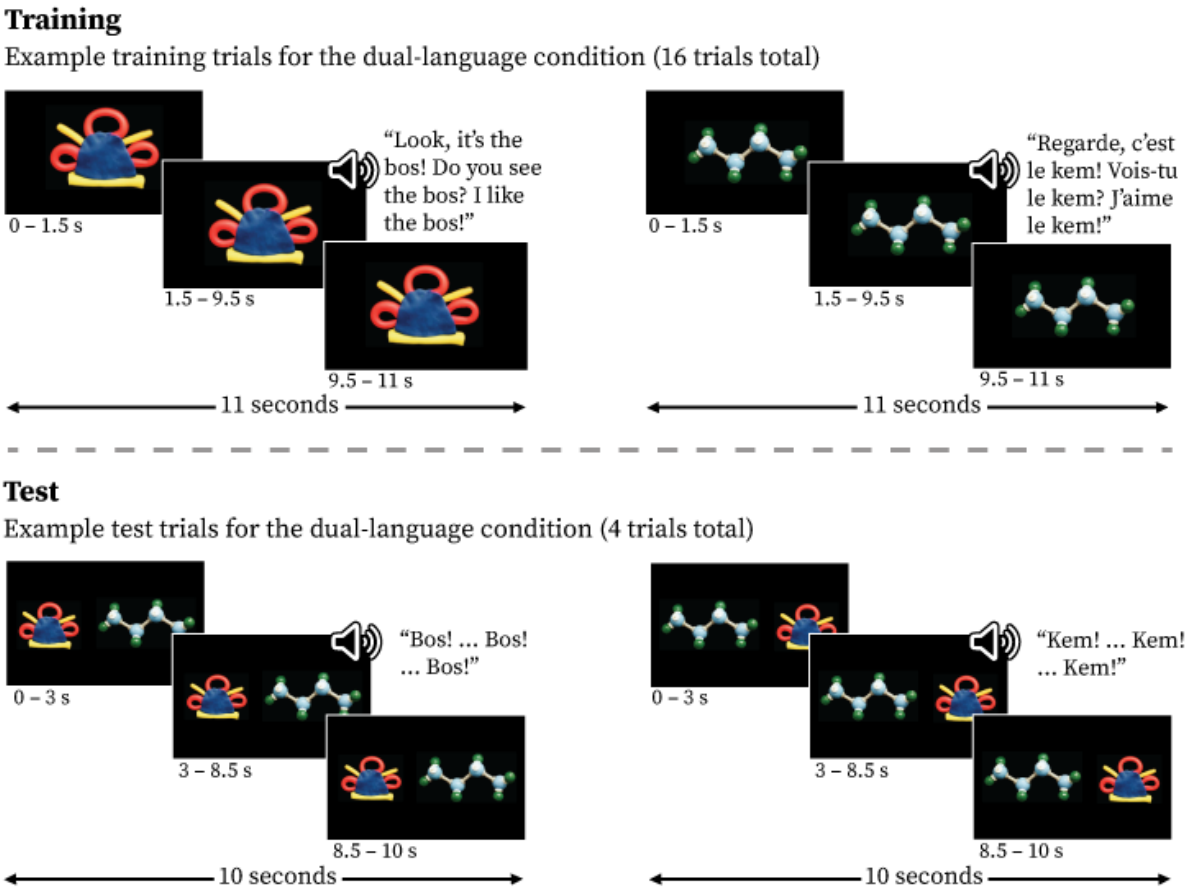


Figure 1. ***Examples of the trial sequence for the training and test phases of the dual-language condition. The single-language condition was identical except all carrier phrases were in the same language.***

Design

Two experimental conditions were developed for the current study: a single-language condition and a dual-language condition. In the single-language condition both objects were labeled in the same language (either English or French) during training. In the dual-language condition, one object was labeled in English and the other object was labeled in French during training. Regardless of condition, each object was labeled 3 times per trial across 8 trials for a total of 24 labeling events per object. Infants thus encountered a total of 16 training trials, presented in one of 8 pseudo-random orders with the constraint that the same word was not encountered for more than two consecutive trials. Orders counterbalanced which word was encountered first, and for the dual-language

condition, the pairing of word and language (e.g., whether *bos* was presented in English versus French sentences).

The test phase for all orders and both conditions presented the words *kem* and *bos* in an alternating fashion—never repeating a word or the side the target image appeared on twice in a row. This was to avoid infants developing a side strategy (e.g., if the target appeared on the right twice in a row, infants could anticipate that on the third test trial the target would yet again appear on the right). Test trials were counterbalanced such that for every order where the training phase ended with *kem*, the corresponding test phase started with *bos*, and vice-versa (for test orders see https://osf.io/g6nrv). There were four test trials in all orders, two for *kem* and two for *bos*.

Infants participated in one of 7 different experiments (see Table 1), defined by the experimental condition they completed (single-language versus dual-language) and their own language background. Based on these two factors, we coded a derived variable called familiarity, which related to infants' level of exposure to the sentence frame language and had two possible values: most familiar and least familiar. Note that this variable describes familiarity with the sentence frame languages only—a trial coded as most familiar means the most familiar of English and French, not necessarily the language an infant is most exposed to overall. For example, a Spanish–French bilingual with 70% exposure to Spanish and 30% exposure to French would have French sentence frames coded as 'most familiar' and English sentence frames as 'least familiar', since out of the two sentence frame languages, they have more familiarity with French than English.

In the dual-language condition (Experiments 1–4), infants encountered one word in English and the other in French sentence frames. French–English bilinguals were familiar with both languages, so the word encountered in their dominant language was coded as most familiar, and the one encountered in their non-dominant language was coded as least familiar (Experiment 1). Bilinguals exposed to English or French and another language were familiar with one of the sentence frame languages (either English or French, but not both); in some cases the most familiar sentence frame language was the infants' dominant language (Experiment 2), and in other cases it was the infants' non-dominant language (Experiment 3). As monolinguals were also familiar with only one of the sentence frame languages, this language was coded as most familiar (Experiment 4). In the single-language condition (Experiments 5–7), both novel words were encountered in the same sentence frame language, thus all trials had the same level of familiarity to each infant. The bilinguals tested in the single-language condition were all French–English bilinguals and were purposefully tested with stimuli in their dominant language, thus all sentence frames were coded as most familiar (Experiment 5). Familiarity was

coded as most familiar for monolinguals tested with native language sentence frames (Experiment 6), and least familiar for monolinguals tested with sentence frames in the other language (which was foreign to them; Experiment 7).

Under the updated analytic approach, we included percent of exposure to the sentence frame language as a continuous version of the categorical familiarity variable. For example, on trials where the novel word was presented in an English sentence frame, an English monolingual with no exposure to any other language would have an exposure score of 100, a French monolingual with no exposure to English would have a score of 0, a French–English bilingual would have a score of 25 (as one possible value, if they were exposed to English 25% of the time), and a French–Arabic bilingual with no exposure to English would have a score of 0. Thus, higher exposure scores indicate more familiarity with the sentence frame language.

Procedure

A trained research assistant greeted and briefed the parents. Parents then signed the consent form and filled out three questionnaires. The first questionnaire gathered basic demographic information (i.e., infants' general health, birth weight, weeks of gestation and socioeconomic status of the family). The second questionnaire was a detailed interview about the infant's language background starting from birth, using the Language Exposure Questionnaire (LEQ; Bosch & Sebastián-Gallés, 2001) with the Multilingual Approach to Parent Language Estimates (MAPLE; Byers-Heinlein et al., 2019). The third questionnaire (the MacArthur-Bates Communicative Development Inventories: Words and Gestures; Fenson et al., 2007) gathered data on the infant's vocabulary knowledge.

Next, the infant and parent were brought to a sound-attenuated room. The infant sat on the parent's lap in a chair approximately 60 cm away from a Tobii T60 XL eye-tracker, which recorded participants' gaze at 60 Hz. Tobii Studio software was used to display the stimuli on a 24" monitor. Parents were given darkened glasses and headphones playing music, and were instructed not to interact with the child to avoid influencing the infant's responses. Following a 5-point eye-tracking calibration, the experiment started with a 10-second pre-familiarization trial, which consisted of a spinning pinwheel accompanied by a sound. Next, infants saw 16 training trials (8 for *kem* and 8 for *bos*) followed by 4 test trials (2 for *bos* and 2 for *kem*). Between each trial, infants saw an attention-getter (a circle stretching vertically and then horizontally while changing colors) to direct their attention back towards the center of the screen. The experiment ended with the presentation of the spinning pinwheel, and in total lasted approximately 5 minutes.

Results

Traditional Approach

The original experimental design was to conduct a series of individual study conditions with small samples (target n ~ 16) that varied the language(s) of the stimuli (i.e., single-language or dual-language) and the population tested (i.e., monolingual, bilingual English–French, bilingual English/French and another language), and 7 of many possible study conditions (Table 1) were ultimately run.

For the dual-language condition (Experiments 1, 2, 3 and 4), we conducted a preliminary series of paired sample *t*-tests to see if infants preferred the most familiar sentence frame language over the other during training (Table 3). This was to ensure that any differences at test would not be due to differential attention during training. We found no statistically significant differences between groups. However, we found a medium-to-large effect size (Cohen's d = -.64) for bilingual infants dominant in French or English with another second language, who looked longer to training trials in their most familiar language. Given that this was only observed in one of the seven studies and was not statistically significant (even prior to correction for multiple comparisons), this effect is unlikely to be meaningful.

Table 3. ***Total Looking Time in Seconds during Training, Dual-Language Condition***

Language Group	Experiment Number	Most Familiar Language Mean *(SD)*	Least Familiar Language Mean *(SD)*	*t*-test
Bilingual French/English (n = 17)	1	40.27 *(20.88)*	41.72 *(15.94)*	$t(16) = .56$, $p = .583$, $d = .14$
Bilingual Dominant in English/ French and L2 Other (n = 10)	2	51.14 *(12.67)*	42.44 *(14.22)*	$t(9) = -2.02$, $p = .074$, $d = -.64$
Bilingual Dominant in Other Language and L2 English/French (n = 14)	3	48.23 *(16.35)*	46.75 *(16.03)*	$t(13) = -.71$, $p = .490$, $d = -.19$
Monolingual (n = 18)	4	32.65 *(20.25)*	31.69 *(17.41)*	$t(17) = -.34$, $p = .739$, $d = -.08$

Note. L2 refers to infants' non-dominant language.

Preliminary analyses also indicated a slight pre-naming preference for looking at the *kem* object in the period of time before the onset of any utterance during the test phase. A *t*-test comparing the proportion looking to each object visible on screen before the onset of the auditory stimulus during test trials (0–3000 ms) showed a statistically significant preference for the *kem* object (*kem M* = .55, *SD* = .15; *bos M* = .45, *SD* = .15), $t(108) = -3.57$, $p = <.001$, $d = -.34$). To account for this difference, we conducted our main analyses using a preference-corrected dependent variable by subtracting each participant's own pre-naming preference for each object from their proportion looking to that target object. This created a variable where a score of zero would indicate no difference between an infant's looking on a given trial and their pre-naming preference for that object, a score greater than zero would indicate more looking to the target object than their pre-naming preference for that object, and a score less than zero would indicate less looking to the target object than their pre-naming preference for that object. Statistical comparisons were then made against zero instead of 50% chance[5].

Following da Estrela and Byers-Heinlein (2016), only infants with at least one data point for each word (i.e., one for *kem* and one for *bos*) were included in the analyses. A series of *t*-tests revealed that only the bilingual English–French and monolingual infants in the dual-language condition (Experiments 1 and 4) looked at the correct object above chance, but only when the novel word was presented in the least familiar sentence frame language during training (Experiment 1: $M = .07$, $SD = .1$, $t(16) = 2.81$, $p = .012$, $d = .68$; Experiment 4: $M = .15.$, $SD = .21$, $t(17) = 2.94$, $p = .009.$, $d = .69$; see Figures 2 and 3; Table 4). This result was surprising, especially for the monolingual group, given that infants were completely unfamiliar with the sentence frame language. We expected this to be the most challenging context for word learning.

To investigate whether the small sample sizes per group were masking an overall effect, we also performed a *t*-test comparing proportion looking minus infants' pre-naming preference for the target object to zero pooling data from all experiments. This test showed that, on average, infants did look slightly above chance during the test phase ($M = .04$, $SD = .21$), $t(219) = 2.88$, $p = .004$, $d = .19$). Further exploratory analyses suggested that this effect was driven by correct looking to the *bos* object when it was labeled ($M = .05$, $SD = .20$, $t(110) = 2.85$, $p = .005$, $d = .27$), but not the *kem* object ($M = .03$, $SD = .21$), $t(110) = 1.26$, $p = .209$, $d = .12$), above chance levels. Thus, when data were

[5] For transparency, we note that the baseline preference for the *kem* object was discovered during the review process. Earlier versions of the manuscript conducted analyses with comparisons to 50% chance. Results were somewhat similar, except that without the baseline correction we found no evidence from either the traditional or updated analyses that infants learned either of the two words.

pooled, we found possible evidence for learning one of the words, but limited to no evidence for learning the other.

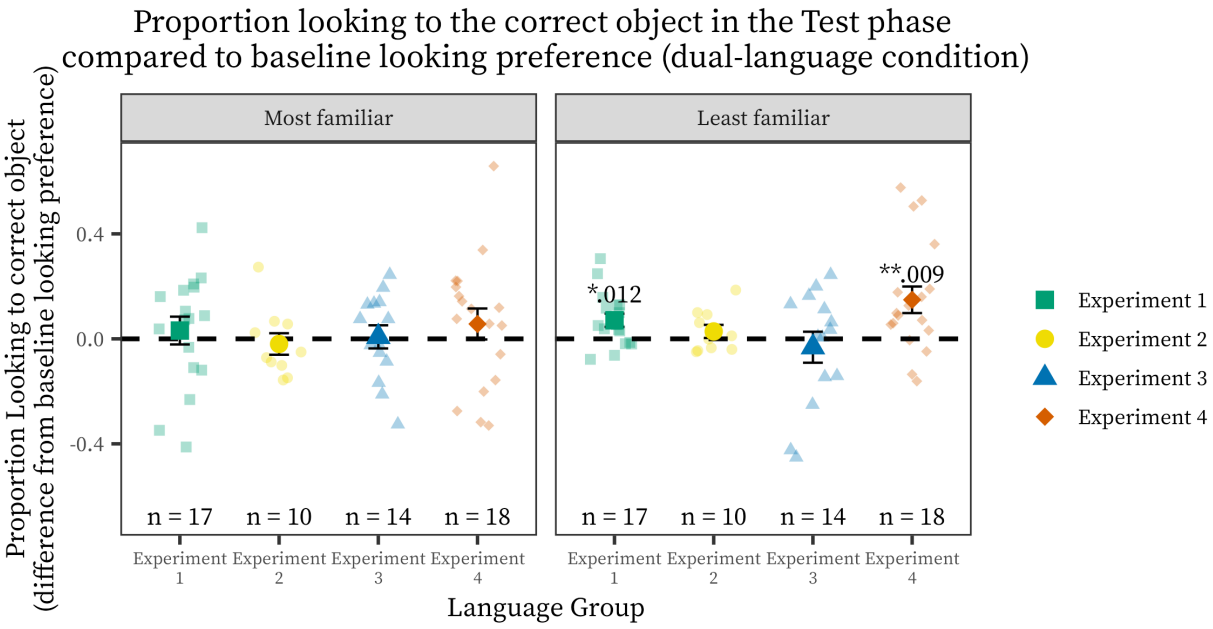


Figure 2. ***Graphs showing proportion looking to the correct object (difference from baseline looking preference) by group in the dual-language condition and standard errors. Same-colour shapes represent an experimental language group. The teal squares represent Experiment 1 (English–French bilinguals). The yellow circles represent Experiment 2 (bilinguals whose first language is English or French with a second language that is not English or French). The blue triangles represent Experiment 3 (bilinguals whose first language is not English or French with English or French as their second language). The orange diamonds represent Experiment 4 (English or French monolinguals). Data are faceted by infants' familiarity with the sentence frame language. Large shapes represent the mean, small shapes represent individual data points, error bars represent the Standard Error, and the dotted line represents no difference from baseline looking preference. The number of participants per mean is indicated with "n =". * p < .05, ** p < .01***

Table 4. ***t-test Results and Means by Group and Condition for the Traditional Analytic Approach***

Language Group	*n*	Sex	Exp. #	Familiarity	Mean	*SD*	*t*	*p*	*df*	*d*
Dual-Language Condition										
Bilingual English–French	17	7 F	1	Most Familiar (Dominant)	0.03	0.22	0.60	0.559	16	0.14
				Least Familiar (Non-Dominant)	0.07	0.1	2.81	**0.012***	16	0.68
Bilingual Dominant in EN/FR and L2 Other language	10	6 F	2	Most Familiar (Dominant)	-0.02	0.13	-0.48	0.644	9	-0.15
				Least Familiar (Foreign)	0.03	0.08	1.13	0.289	9	0.36
Bilingual Dominant in Other language and L2 EN/FR	14	8 F	3	Most Familiar (Non-Dominant)	0.01	0.16	0.17	0.867	13	0.05
				Least Familiar (Foreign)	-0.03	0.22	-0.54	0.598	13	-0.14
Monolingual	18	8 F	4	Most Familiar (Native)	0.06	0.25	0.96	0.351	17	0.23
				Least Familiar (Foreign)	0.15	0.21	2.94	**0.009****	17	0.69
Single-Language Condition										
Bilingual EN/FR	19	12 F	5	Most Familiar (Dominant–*Kem*)	0.02	0.15	0.68	0.503	18	0.16
				Most Familiar (Dominant–*Bos*)	0.05	0.17	1.31	0.206	18	0.30
Monolingual	16	9 F	6	Most Familiar (Native–*Kem*)	-0.01	0.36	-0.07	0.944	15	-0.02
				Most Familiar (Native–*Bos*)	0.04	0.28	0.59	0.566	15	0.15
Monolingual	16	7 F	7	Least Familiar (Foreign–*Kem*)	0.06	0.19	1.32	0.206	15	0.33
				Least Familiar (Foreign–*Bos*)	0.05	0.2	0.92	0.371	15	0.23

Note. L2 refers to infants' non-dominant language. F refers to number of females. EN = English, FR = French. Exp. = Experiment number.

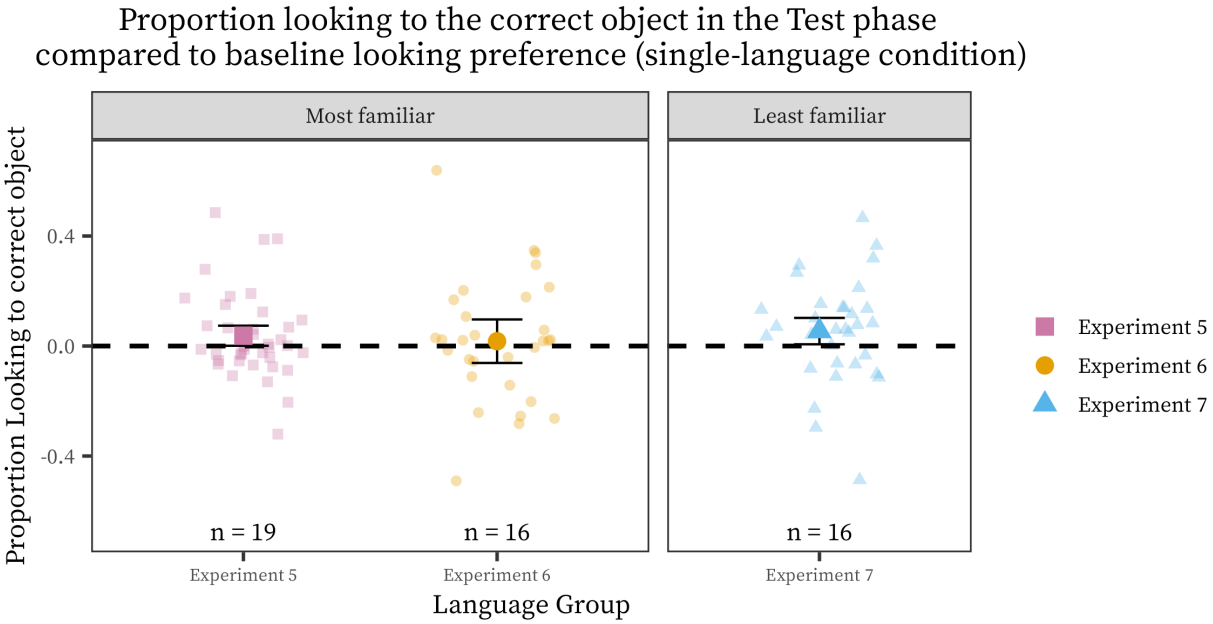


Figure 3. ***Graphs showing proportion looking to the correct object (difference from baseline looking preference) by group in the single-language condition. The purple squares represent Experiment 5 (English–French bilinguals), the orange circles represent Experiment 6 (English or French monolinguals tested in their native language), and the light blue triangles represent Experiment 7 (English or French monolinguals tested in the language they do not know). Large shapes represent the mean, small shapes represent individual data points, error bars represent the Standard Error, and the dotted line represents no difference from baseline looking preference. The number of participants per mean is indicated with "n =". Data are faceted by infants' familiarity with each sentence frame language.***

Updated Approach

Our traditional approach largely tested the performance of small groups of participants against chance level, following the relevant literature at the time the study was designed. More recent discussions on the reproducibility and reliability of psychological science highlight the need for more sensitive analytical approaches that take into consideration the structure of the data (e.g., repeated measures) and that have an appropriate sample size (Bergmann et al., 2018; Oakes, 2017). One well-accepted approach is mixed-effects models (Dixon, 2008). These models have several advantages over traditional methods such as ANOVAs or multiple *t*-tests on different groups. For instance, they can account for the relationship between continuous outcomes (e.g., looking time to the target) and

continuous predictors (e.g., language exposure, vocabulary size), which are modeled as fixed effects. They can also account for systematic variability arising from data being grouped (e.g., repeated measures within participants or items), which are modeled as random effects. Furthermore, by modelling fine-grained data (e.g., trial-level data rather than condition averages), these models have greater statistical power and better handling of missing data, even for unbalanced datasets (Baayen et al., 2008; Bates et al., 2015). To harness the richness of our eye-tracking data, we fitted linear mixed-effects models to investigate infant word learning, using the lme4 package for *R* (Bates et al., 2015). All data and scripts are available at https://osf.io/upy7f.

For this analytical approach, we used a larger sample (N = 148; see Participants for details). Although larger, we must note that this sample was highly heterogeneous, with infants from diverse linguistic backgrounds (Table 2). We tested whether in this larger sample infants showed word learning and the influence of covariates such as familiarity with sentence-frame language, receptive vocabulary size, and total looking time to objects during the training phase. Given that our Traditional Analysis revealed a preference for *bos* over *kem*, we included target words as a random effect in the model, which would allow us to test the effects of our predictors of interest on word learning while controlling for any differences in looking between the two target objects.

The dependent variable for mixed-models was the proportion of looking time to the labeled object in each trial minus the chance level (.5), so that the intercept would capture overall word learning different from chance. First, we fit an intercept-only model to examine infants' mean accuracy before exploring potential moderators of performance (Table 5). Next, we explored the effects of three continuous variables on learning: the percent of exposure to the sentence frame language, infants' receptive vocabulary size in the sentence frame language, and the total looking time to the objects during the training phase (Table 6). Percentage of exposure to the sentence frame language and vocabulary size allowed us to further explore if or how our participants' language background guided learning. Total looking time to the objects during training allowed us to investigate if participants who were more or less attentive during training would show differences in learning during the test phase. We also ran models on the conditions separately (i.e., one model for the dual-language condition and one for the single-language condition), to see if combining them might be masking some effects. However, there were no additional effects, so these models are not reported here (see Supplemental Materials, Tables S6 to S8, available at https://osf.io/upy7f).

We attempted to fit a maximal random effects structure to our models that included the novel words (*kem* and *bos*) as random slopes and participants as random intercepts (Barr

et al., 2013). These models had a singular fit. We then attempted to include the novel words and participants as separate random intercepts. Once again, the models had a singular fit and a closer inspection indicated that there was not enough variability between participants to be included as random intercepts. We then simplified the models to include only the target words as random intercepts. These models converged without a singularity warning and respected the assumptions of normality (see https://osf.io/upy7f for details).

Results are displayed in Tables 5 and 6, and Figure 4. Overall, our reanalysis with this updated approach and the larger sample size confirmed the pattern found in the traditional analyses: there was no evidence of overall word learning while controlling for the difference in looking between *kem* and *bos*, and further, there were no significant relationships between the proportion of looking to the target and (a) exposure to the sentence frame languages, (b) receptive vocabulary size in the sentence frame languages, or (c) the total looking time to the objects during training. Estimates were close to zero for the intercept as well as for all predictors. This means that none of our variables of interest predicted the proportion of infants' looking at the labeled objects (Table 6). Furthermore, our approximate effect size, calculated from the intercept-only model using Brysbaert and Stevens' (2018) approach, was very small (d = 0.09).

Table 5. ***Fixed and Random Effects for the Intercept-Only Model [proportion of looking time - .5 ~ 1 + (1 \| target word)]***

Predictors	Estimates	95% Confidence Interval	*p*
Intercept	0.03	-0.06 – 0.11	0.544
Random Effects			
σ^2			0.06
$\tau_{00\ target\ word}$			0.00
ICC			0.05
Observations	476		
Marginal R^2 / Conditional R^2	0.000 / 0.049		

Table 6. ***Fixed and Random Effects for the Pruned Model with Exposure to Sentence Frame Language, Vocabulary size, and Total Looking Time during Training as Predictors of Looking to the Labeled Object [proportion of looking time - .5 ~ exposure + vocabulary + total looking time during training + (1 | target word)]***

Predictors	Estimates	95% Confidence Interval	*p*
Intercept	0.05	-0.06 – 0.16	0.334
Exposure to sentence frame language	-0.00	-0.00 – 0.00	0.557
Receptive vocabulary for sentence frame language	0.00	-0.00 – 0.00	0.409
Total looking time during training	-0.00	-0.00 – 0.00	0.369
Random Effects			
σ^2			0.06
$\tau_{00\ target\ word}$			0.00
ICC			0.05
Observations	476		
Marginal R^2 / Conditional R^2	0.003 / 0.052		

To follow up on the finding from the traditional analysis where we found some evidence of learning on *bos* test trials (but not on *kem* test trials), we attempted to fit a model with target word as a fixed effect in addition to our other predictors as fixed effects. Again, models were singular when participants were included as a random effect. Thus, we ran a multiple linear model with these data using the preference difference score as the dependent variable to account for baseline differences in looking toward the two objects. We again found evidence that performance was better for *bos* trials than *kem* trials (see Table 7), after infants' pre-naming baseline looking preferences were accounted for ($\beta 0 = .07$, $p = .031$). However, no other predictors were significant and the model overall explained very little variance in the data ($R^2 = .007$, $F(4,469) = .877$, $p = .477$).

Overall, depending on the model, we found either little evidence of word learning or some evidence of learning one but not both words. Our models also provided little to no account of the observed variance.

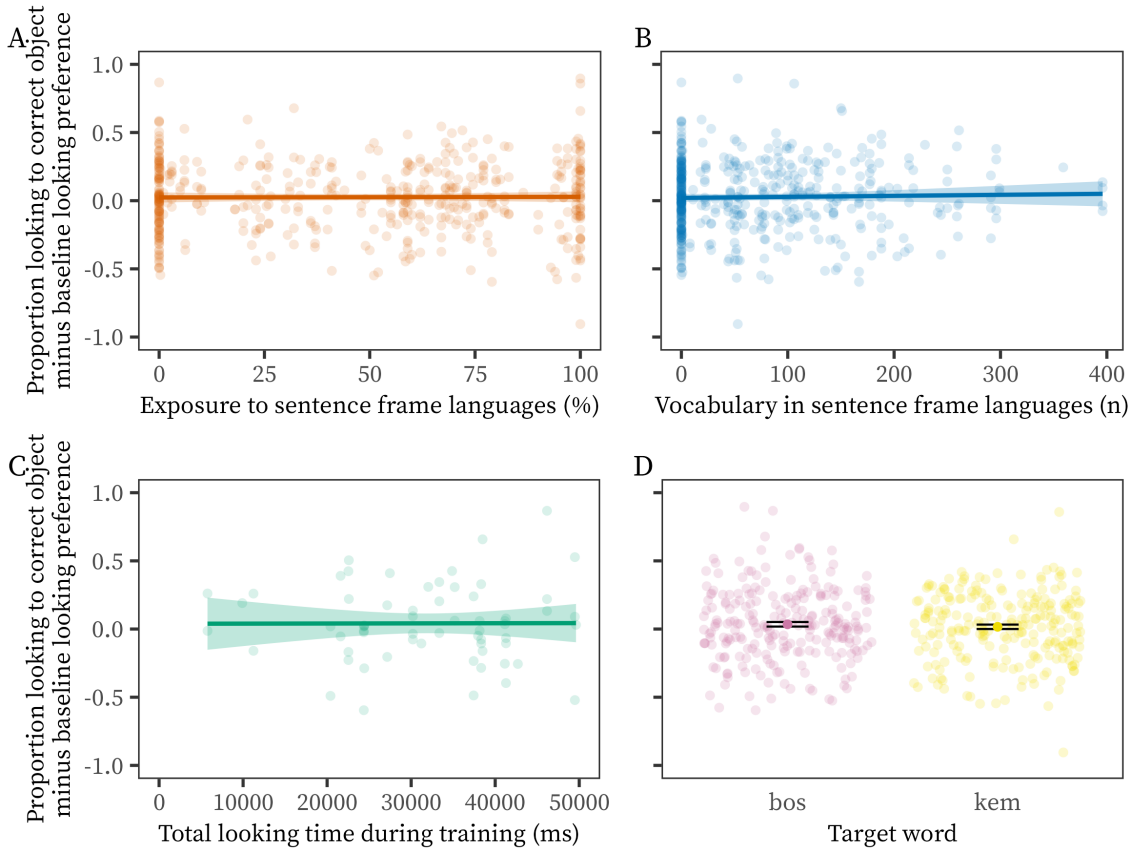


Figure 4. ***Proportion of looking at the correct object as a function of (A) percentage of exposure to the sentence frame languages, (B) vocabulary size (number of words comprehended) in the sentence frame languages, (C) total looking time (ms) during the training phase, and (D) target word. Regression lines, standard errors, and all data points are plotted. Note that chance is 0.***

Table 7. ***Multiple linear regression results using difference score as the criterion***

Predictors	*Estimates*	95% Confidence Interval	*p*
Intercept	0.08	0.01 – 0.15	**0.048**
Target word [*kem*]	-0.02	-0.06 – 0.03	0.439
Exposure to sentence frame language	-0.00	-0.00 – 0.00	0.792
Receptive vocabulary for sentence frame language	0.00	-0.00 – 0.00	0.674
Total looking time during training	-0.00	-0.00 – 0.00	0.165
Observations	476		
R^2 / R^2 adjusted	0.006 / -0.002		

General Discussion

The present study investigated word learning in 14-month-olds from different language backgrounds using a preferential looking paradigm. Following prior research (Fennell & Waxman, 2010; Fernald & Hurtado, 2006; Havy et al., 2016) we assumed that the use of sentence frames would support word learning in infants, and that infants would readily learn the two words that they encountered during the training phase. Moreover, we predicted that language familiarity would play a key role in word learning, with infants showing better learning of word–object associations when they had greater familiarity with the sentence frame language.

First, and quite surprisingly we found only limited evidence for successful word learning in this paradigm. Out of 7 *t*-tests conducted in our traditional approach, only two showed performance that was statistically above chance overall, which we interpret as possible false positives, although given the small samples (n = 10–18/group) false negatives in the other experiments are also possible. Moreover, in our updated approach, which used a larger dataset (N = 148) and had greater statistical power (reducing the chances of both Type I and Type II error), mixed effects models found no evidence of an effect of amount of exposure to the sentence frame language, vocabulary, or attention during the training phase on word learning. By contrast, when data were pooled without including a random effect for item (via *t*-tests and linear regressions), there was some evidence that infants learned one, but not both words. Specifically, when baseline looking preferences were

taken into account, there was evidence that infants learned "*bos*" but not "*kem*". We note that successful learning of both of these nonsense words has been previously reported in the literature (Fennell & Byers-Heinlein, 2014; Mattock et al., 2010), making it unlikely that this pattern was driven by our particular choice of stimuli. Overall, evidence for successful word learning in this study was inconsistent.

With respect to familiarity effects, again there was only limited and weak evidence in a direction contrary to hypotheses. Specifically, when traditional analyses were conducted (via separate *t*-tests on data from small groups of infants), two groups of infants showed evidence of learning words presented in frames that were in their least familiar language, but none showed evidence of learning words presented in frames that were in their most familiar language. Again, we note that these analyses had limited statistical power. However, in the updated linear mixed-effects models, which measured familiarity continuously, we did not find an effect of familiarity.

Overall, we believe that the most appropriate interpretation of our results is that word learning in the lab using this paradigm can be challenging for some infants, even with supporting sentence frames. Our findings are unexpected and contrast with previous studies that have reported successful word learning for monolingual 14-month-olds using isolated words (Graf Estes & Hay, 2015; Werker et al., 1998; Yin & Csibra, 2015) and sentence frames (da Estrela & Byers-Heinlein, 2016; Fennell & Waxman, 2010). Importantly, our task was designed to be easy and conducive to word learning. To this end, we used sentence frames which were meant to provide further linguistic cues and presented the target words in a sentence-final position to increase their salience (e.g., Fennell & Byers-Heinlein, 2014; Fernald & Hurtado, 2006). In addition, each word was repeated multiple times during training (3 times per trial for 8 trials, for a total of 24 exposures to each word–object pairing) and we taught infants only two novel words to reduce their cognitive load. Even so, neither monolingual nor bilingual infants showed evidence of learning both words, even the word–object pairs presented in the sentence frame language that was most familiar to them.

Although our experiment was designed to provide a facilitative word learning opportunity for infants, it is possible that the task was simply too taxing. We used consistent word–object pairings that have been used successfully in previous studies of word learning (Werker et al. 1998; Werker et al., 2002; Fennell et al., 2007), but it is possible that these stimuli were suboptimal[6]. One crucial difference between our study

[6] For example, the pairings might have violated sound symbolic associations (e.g., Sidhu & Pexman, 2018). We thank an anonymous reviewer for raising this point.

and previous studies that have successfully shown word learning with 14-month-olds (e.g., Werker et al., 1998) is that our study presented infants with a fixed number of training trials rather than presenting training trials according to a habituation criterion (as in the Switch task), which may make our task less effective as it did not adapt to each infant's learning (Yoshida et al., 2009). It could also be the case that infants required additional familiarization with the task structure (e.g., familiar-word trials presented before training, where a known word is associated with a known object to cue the task, see Fennell & Waxman, 2010 and May & Werker, 2014). However, this interpretation contrasts with reports in the published literature. For example, Schafer and Plunkett (1998) reported successful word learning after 12 presentations of each of 2 novel word–referent pairs in 15-month-olds using a similar paradigm to that implemented in our study (though they also presented familiar-word trials between the novel word trials). It is also possible that, rather than presenting infants with too few training trials, we presented them with too many, ultimately leading to boredom and disengagement from the task. This interpretation is supported by the high levels of attrition we observed in our task, a point that we return later in this section. Overall, the optimal amount of exposure to novel words in lab word learning tasks remains unclear.

It is also possible that sentence frames made our task more challenging, contrary to our intentions. We used sentence frames following prior research with monolingual and bilingual infants showing that they have a facilitative effect (e.g., Fennell & Waxman, 2010 in 14-month-olds; Fennell & Byers-Heinlein, 2014 in 17-month-olds; Fernald & Hurtado, 2006 in 18-month-olds). Thus, we expected that sentence frames would support word learning, particularly for bilingual infants, since this additional information might help them identify the language in which a novel word is presented. Yet, this did not appear to be the case. Similarly, it is possible that using isolated words during testing might have made the task more challenging, since during training sentence frames were used. Future studies could compare experimental conditions that vary on the use of isolated words versus sentence frames (e.g., Morini & Newman, 2019), to disentangle the effect that additional linguistic information has on early word learning.

Another possible explanation is that infants did successfully learn both words presented during training trials, but our test phase was not sufficiently sensitive to detect this learning. It could be that the 4 test trials included in our study (2 per novel word) were not enough to robustly detect learning, especially because some infants did not provide valid data for both words during the test phase. Prior studies using a preferential looking paradigm reported successful word learning when infants were tested with 4–8 novel word test trials per condition (e.g., Chen et al., 2020; Schafer & Plunkett, 1998; Tan & Schafer, 2005; Yoshida et al., 2009), although studies using the Switch word learning

paradigm have often used only two test trials (see data compiled by Tsui et al., 2019). Increasing the number of test trials per infant might increase the chances of capturing learning in this hard-to-test population, and would most likely generate a better representation of infants' true response to the task, thus decreasing noise and increasing statistical power (DeBolt et al., 2020).

Moreover, we selected the preferential looking task based on extant literature suggesting that it might be more sensitive to detect word learning than other paradigms such as the Switch task (Yoshida et al., 2009). However, many studies reporting successful word learning in infants have used the Switch task (see Tsui et al., 2019 for a meta-analysis), and it may be that the Switch task is in fact more sensitive, or at least more forgiving when infants have only learned one of two words. In the Switch task at least two novel words are paired with two referents (word A with object A, word B with object B). At test, some trials show the label and referent that were previously paired (A with A; Same trials) and some trials show a label with the other referent (A with B; Switch trials). In this paradigm, infants only need to associate one word–referent pair to recognize a word–object violation. If infants learn that word A should be associated with object A, they should be able to detect the violation when word A is paired with object B. However, in our preferential looking paradigm, infants had to correctly identify both word–object pairings to show learning of each word. Moreover, it may be that detecting a pairing violation (dishabituating in the Switch task) can potentially be accomplished with weaker knowledge than looking towards a correct referent in a preferential looking paradigm. Tsui et al.'s (2019) meta-analysis reported an average effect of Cohen's d = 0.33, 95% CI [0.03, 0.63] in comparable studies using the Switch task (i.e., 14-month-olds learning dissimilar-sounding words), which was moderate and much larger than the approximate d = 0.09 we observed in our own data (Brysbaert & Stevens, 2018). Nonetheless, little work has compared infants' performance in the Switch task to a preferential looking test using the same learning task (although see Yoshida et al., 2009), and thus it remains an open methodological question which tasks are most sensitive for testing infant word learning. Developing maximally sensitive and reliable tasks should be a priority for research on infant word learning.

Another well-documented possibility is that sampling and measurement error in the context of small samples can lead to highly variable, and unreliable, effect-size estimates (Brysbaert, 2021; Lindsay, 2020; Oakes, 2017). For instance, underpowered studies can lead to exaggerated effect size estimates that, combined with publication bias favouring positive results, might end up published, whereas null results with similar sample sizes end up in the file drawer (Rosenthal, 1979). As mentioned in the Introduction, our per group sample size was chosen back in 2012, following sample sizes from other studies in

the field (e.g., Fennell & Werker, 2003; Mattock et al., 2010), and after 7 years of testing infants, we were not able to achieve our (small) target sample in all groups. In retrospect, we acknowledge that our original experimental plan was both overly ambitious and underpowered. Even when these small groups were combined in our updated approach, the sample was very heterogeneous, limiting our explanatory scope. At the same time, given our large overall sample, we would have expected to find statistically reliable learning of both words, even if there were some moderators of an overall positive effect size. However, our mixed effects models explored three different variables – percent exposure to the sentence frame language, receptive vocabulary in the sentence frame language, and attention during training – and found no effects (estimated effect size of d = 0.09). In fact, it was surprising that neither percentage exposure nor vocabulary size modulated performance in this task, given prior studies reporting the influence of these variables in word learning (e.g., Bion et al., 2013; Werker et al., 2002).

Despite these unexpected and mixed results, we believe that there is value in sharing our study, as it shows some of the drawbacks to using traditional methodologies and conventional sample sizes. Open science practices centered around transparency and collaboration, combined with more advanced statistical analyses, have an enormous potential to inform future studies on infant word learning. By planning adequate sample sizes (using a-priori power analyses and simulations), pre-registering analytical pipelines, and sharing materials, data, and research reports, we can work toward more reliable findings in the field. For instance, readers can use our openly shared materials, data, and analysis scripts (open repository: https://osf.io/upy7f/) to both reproduce our methodological and analytical decisions and build on them when designing future investigations on the topic.

Another important issue our study faced is the reduction of our initial sample size. Though we tested 288 14-month-old infants, after implementing our exclusion criteria we lost 62% of our participants for the traditional approach and 49% for the updated one. A large proportion of our exclusions (23% for the traditional approach) were related to infants' language background, which can be a particular challenge of studies with bilingual populations. Within the other excluded infants, the largest reason for exclusion was fussiness and inattention (15%), a major issue in infant research. In our updated analytical approach, we were able to include 38 additional participants who had been excluded using the traditional analytic approach. Including these additional infants did not change the pattern of results that we observed. Moreover, our attrition rates, while high, are within the range reported in previous studies including infants of similar ages (e.g., 26% exclusion rate in Experiment 1 and 32% in Experiment 2 in Graf Estes et al., 2007; 44% exclusion rate in Yu & Smith, 2011, 35% exclusion rate in Escudero &

Kalashnikova, 2020). While high attrition can reduce power, our sample size was still large overall.

One way to achieve larger sample sizes and more robust results is with collaborations between different research labs. When participants are recruited in multiple locations, it is easier to obtain larger samples, and the results are also more generalizable. Although some researchers may find it more challenging than others to conduct large studies on their own or to engage in large-scale collaborations, it is important to consider the value of carrying out research that may not be sufficiently powered in the first place (Brysbaert, 2021; Oakes, 2017). In recent years, more opportunities to take part in such large-scale collaborations have become available, and often do not require extensive resources to join (e.g., Byers-Heinlein et al., 2020; Frank et al., 2017; ManyBabies Consortium, 2020). Similarly, open science practices such as open sharing of stimuli and protocols between researchers can be useful in identifying procedures and materials (e.g., novel objects, number of trials) that are better tolerated by infants at specific ages, reducing fussiness and participant loss.

There are several other explanations for weak or null results that we also considered, but found unlikely[7]. First, it has been proposed that null results in some infant looking time studies could be due to some infants showing a familiarity preference, and others a novelty preference, which could average out to a null result (e.g., DePaolis et al., 2016). However, this line of reasoning does not clearly apply to preferential looking paradigms like ours, where infants are always expected to look towards the labeled target rather than the distractor object. Second, one might ask whether incidental factors such as the room where data were collected, or the particular speaker who recorded our stimuli, contributed to our null results. Our lab has conducted many other studies with positive results in the same space, and using similar procedures for recording stimuli, training research assistants, and testing infants, making it unlikely that these factors would affect this study in particular. In the future, Big Team Science efforts such as ManyBabies might provide insight into whether and how such incidental sources of variation relate to effect sizes (Frank et al., 2017). Third, it is possible that there is an error in our data analysis pipeline. However, this seems unlikely as looking time was gathered via an eyetracker, and the analysis was fully automated in *R* and was double checked. We have provided our materials, raw data, and analysis code on the Open Science Framework such that it can be checked or even further analysed by other researchers, who might come to different conclusions or identify limitations that we did not. We would welcome this type of feedback.

[7] We acknowledge the peer-review process for raising these possibilities.

Regardless of any potential limitations of the current experiment as designed and performed, our results are nonetheless surprising. There is a vast body of published research showing successful word learning in the lab with infants, even with methods and small sample sizes comparable to ours. Taking into account the publication bias for positive results (Carter et al., 2019; Rosenthal, 1979), it is impossible to know how many "unsuccessful" infant word learning studies languish in the file drawer. If they do exist in significant numbers, their absence from the literature may distort the picture of how easily infants learn new words in the lab, and, by consequence, any generalization to the real world outside the lab. Increasingly, journals, editors, and reviewers are recognizing the importance of publishing null results, and researchers are embracing open science practices such as pre-registration and registered reports (Tsuji et al., 2020a). With these efforts, the published literature might present a more accurate picture of the true effects in hard-to-test populations, like young infants. Additionally, developing large-scale collaborations across labs, with greater power and sample diversity, might also contribute to a better characterization of infants' word learning abilities.

Overall, our study raises the possibility that word learning in the lab could in some cases be challenging for 14-month-old bilinguals and monolinguals, despite the presence of sentence frames that could support learning. The case study presented here highlights the need for and value of open science practices to advance our understanding of infant development.

References

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*(4), 390–412. https://doi.org/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*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Bergelson, E. (2020). The comprehension boost in early word learning: Older infants are better learners. *Child Development Perspectives, 14*(3), 142–149. https://doi.org/10.1111/cdep.12373

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences, 109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development, 89*(6), 1996–2009. https://doi.org/10.1111/cdev.13079

Bion, R. A., 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*(1), 39-53. https://doi.org/10.1016/j.cognition.2012.08.008

Bishop, D. V. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research. *Quarterly Journal of Experimental Psychology, 73*(1), 1–19. https://doi.org/10.1177/1747021819886519

Bosch, L., & Sebastián-Gallés, N. (2001). Evidence of early language discrimination abilities in infants from bilingual environments. *Infancy, 2*(1), 29–49. https://doi.org/10.1207/S15327078IN0201_3

Brysbaert, M. (2021). Power considerations in bilingualism research: Time to step up our game. *Bilingualism: Language and Cognition, 24*(5), 813-818. https://doi.org/10.1017/S1366728920000437

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1*(1), 9. https://doi.org/10.5334/joc.10

Burnham, D., Singh, L., Mattock, K., Woo, P. J., & Kalashnikova, M. (2018). Constraints on tone sensitivity in novel word learning by monolingual and bilingual infants: tone properties are more influential than tone familiarity. *Frontiers in Psychology, 8*, 2190. https://doi.org/10.3389/fpsyg.2017.02190

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. https://doi.org/10.1038/nrn3475

Byers-Heinlein, K. (2013). Parental language mixing: Its measurement and the relation of mixed input to young bilingual children's vocabulary size. *Bilingualism: Language and Cognition, 16*(1), 32–48. https://doi.org/10.1017/S1366728912000120

Byers-Heinlein, K. (2015). Methods for studying infant bilingualism. In J. W. Schwieter (Ed.), *The Cambridge Handbook of Bilingual Processing* (pp. 133–154). Cambridge University Press. https://doi.org/10.1017/CBO9781107447257.005

Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Hamlin, J. K., Kline, M., Kominsky, J. F., Kosie, J. E., Lew-Williams, C., Liu, L., Mastroberardino, M., Singh, L., Waddell, C. P. G., Zettersten, M., & Soderstrom, M. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology/Psychologie Canadienne, 61*(4), 349–363. https://doi.org/10.1037/cap0000216

Byers-Heinlein, K., Fennell, C. T., & Werker, J. F. (2013). The development of associative word learning in monolingual and bilingual infants. *Bilingualism: Language and Cognition, 16*(1), 198–205. https://doi.org/10.1017/S1366728912000417

Byers-Heinlein, K., Jardak, A., Fourakis, E., & Lew-Williams, C. (2022). Effects of language mixing on bilingual children's word learning. *Bilingualism: Language and Cognition, 25*(1), 55-69. doi:10.1017/S1366728921000699

Byers-Heinlein, K., Schott, E., Gonzalez-Barrero, A. M., Brouillard, M., Dubé, D., Jardak, A., Laoun-Rubenstein, A., Mastroberardino, M., Morin-Lessard, E., Pour Iliaei, S., Salama-Siroishka, N., & Tamayo, M. P. (2019). MAPLE: A Multilingual Approach to Parent Language Estimates. *Bilingualism: Language and Cognition*, 1–7. https://doi.org/10.1017/S1366728919000282

Byers-Heinlein, K., Tsui, A. S. M., Bergmann, C., Black, A. K., Brown, A., Carbajal, M. J., … & Wermelinger, S. (2021). A multilab study of bilingual infants: Exploring the preference for infant-directed speech. *Advances in Methods and Practices in Psychological Science, 4*(1), 1-30. https://doi.org/10.1177/2515245920974622

Byers-Heinlein, K., & Werker, J. F. (2009). Monolingual, bilingual, trilingual: Infants' language experience influences the development of a word-learning heuristic. *Developmental Science, 12*(5), 815–823. https://doi.org/10.1111/j.1467-7687.2009.00902.x

Canfield, R. L., Smith, E. G., Brezsnyak, M. P., Snow, K. L., Aslin, R. N., Haith, M. M.,

Wass, T. S., & Adler, S. A. (1997). Information processing through the first year of life: A longitudinal study using the visual expectation paradigm. *Monographs of the Society for Research in Child Development, 62*(2), 1-160. https://doi.org/10.1111/j.1540-5834.1997.tb00511.x

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford Child Language Conference, 15*, 17–29.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science, 2*(2), 115–144. https://doi.org/doi.org/10.1177/2515245919847196

Chen, H., Labertonière, D., Cheung, H., & Nazzi, T. (2020). Infant learning of words in a typologically distant nonnative language. *Journal of Child Language, 47*(6), 1276–1287. https://doi.org/10.1017/S0305000920000161

Curtin, S., Fennell, C. T., & Escudero, P. (2009). Weighting of vowel cues explains patterns of word-object associative learning. *Developmental Science, 12*(5), 725–731. https://doi.org/10.1111/j.1467-7687.2009.00814.x

da Estrela, C., & Byers-Heinlein, K. (2016). Vois-tu le kem? Do you see the bos? Foreign word learning at 14 months. *Infancy, 21*(4), 505–521. https://doi.org/10.1111/infa.12126

Dal Ben, R., de Hollanda Souza, D., & Hay, J. F. (2019). Cross-situational word learning: Systematic review and meta-analysis [Manuscript in preparation]. https://doi.org/10.17605/OSF.IO/GU9RB

DeBolt, M. C., Rhemtulla, M., & Oakes, L. M. (2020). Robust data and power in infant research: A case study of the effect of number of infants and number of trials in visual preference procedures. *Infancy, 25*(4), 393–419. https://doi.org/10.1111/infa.12337

DePaolis, R. A., Keren-Portnoy, T., & Vihman, M. (2016). Making sense of infant familiarity and novelty responses to words at lexical onset. *Frontiers in Psychology, 7*, 715. https://doi.org/10.3389/fpsyg.2016.00715

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59*(4), 447–456. https://doi.org/10.1016/j.jml.2007.11.004

Echols, C. H., & Newport, E. L. (1992). The role of stress and position in determining first words. *Language Acquisition, 2*(3), 189–220. https://doi.org/10.1207/s15327817la0203_1

Escudero, P., & Kalashnikova, M. (2020). Infants use phonetic detail in speech perception and word learning when detail is easy to perceive. *Journal of Experimental Child Psychology, 190,* 104714. https://doi.org/10.1016/j.jecp.2019.104714

Fennell, C. T., & Byers-Heinlein, K. (2014). You sound like Mommy: Bilingual and monolingual infants learn words best from speakers typical of their language environments. *International Journal of Behavioral Development, 38*(4), 309–316. https://doi.org/10.1177/0165025414530631

Fennell, C. T., Byers-Heinlein, K., & Werker, J. F. (2007). Using speech sounds to guide word learning: The case of bilingual infants. *Child Development, 78*(5), 1510–1525. https://doi.org/10.1111/j.1467-8624.2007.01080.x

Fennell, C. T., & Waxman, S. R. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child Development, 81*(5), 1376–1383. https://doi.org/10.1111/j.1467-8624.2010.01479.x

Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech, 46*(2–3), 245–264. https://doi.org/10.1177/00238309030460020901

Fenson L., Marchman V. A., Thal D. J., Dale P. S., Reznick S., Bates E. (2007). *MacArthur-Bates communicative development inventories* (2nd ed.). Brookes Publishing.

Fernald, A., & Hurtado, N. (2006). Names in frames: Infants interpret words in sentence frames faster than words in isolation. *Developmental Science, 9*(3), F33–F40. https://doi.org/10.1111/j.1467-7687.2006.00482.x

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and Consistency in Early Language Learning: The Wordbank Project.* MIT Press.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices,

and theory-building. *Infancy, 22*(4), 421–435. https://doi.org/10.1111/infa.12182

Gonzalez-Gomez, N., Poltrock, S., & Nazzi, T. (2013). A "bat" is easier to learn than a "tab": Effects of relative phonotactic frequency on infant word learning. *PLoS One, 8*(3), e59601. https://doi.org/10.1371/journal.pone.0059601

Graf Estes, K. (2014). Learning builds on learning: Infants' use of native language sound patterns to learn words. *Journal of Experimental Child Psychology, 126*, 313–327. https://doi.org/10.1016/j.jecp.2014.05.006

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science, 18*(3), 254-260. https://doi.org/10.1111/j.1467-9280.2007.01885.x

Graf Estes, K., & Hay, J. F. (2015). Flexibility in bilingual infants' word learning. *Child Development, 86*(5), 1371–1385. https://doi.org/10.1111/cdev.12392

Havy, M., Bouchon, C., & Nazzi, T. (2016). Phonetic processing when learning words: The case of bilingual infants. *International Journal of Behavioral Development, 40*(1), 41–52. https://doi.org/10.1177/0165025415570646

Hirsh-Pasek, K., Golinkoff, R. M., & Hollich, G. (2000). An emergentist coalition model for word learning: Mapping words to objects is a product of the interaction of multiple cues. In R. M. Golinkoff, K. Hirsh-Pasek, L. Bloom, L. Smith, A. L. Woodward, N. Akhtar, M. Tomasello, & G. Hollich (Eds.), *Becoming a word learner: A debate on lexical acquisition* (pp. 136–164). Oxford University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kalashnikova, M., Escudero, P., & Kidd, E. (2018). The development of fast-mapping and novel word retention strategies in monolingual and bilingual infants. *Developmental Science, 21*(6), e12674. https://doi.org/10.1111/desc.12674

Kalashnikova, M., Mattock, K., & Monaghan, P. (2015). The effects of linguistic

experience on the flexible use of mutual exclusivity in word learning. *Bilingualism: Language and Cognition, 18*(4), 626-638. https://doi.org/10.1017/S1366728914000364

Kandhadai, P., Hall, D. G., & Werker, J. F. (2017). Second label learning in bilingual and monolingual infants. *Developmental Science, 20*(1), e12429. https://doi.org/10.1111/desc.12429

Kannass, K. N., & Oakes, L. M. (2008). The development of attention and its relations to language in infancy and toddlerhood. *Journal of Cognition and Development, 9*(2), 222–246. https://doi. org/10.1080/15248370802022696

Kremin, L. V., & Byers-Heinlein, K. (2021). Why not both? Rethinking categorical and continuous approaches to bilingualism. *International Journal of Bilingualism, 25*(6), 1560-1575. https://doi.org/10.31234/osf.io/nkvap

Kremin, L. V., Alves, J., Orena, A. J., Polka, L., & Byers-Heinlein, K. (2022). Code-switching in parents' everyday speech to bilingual infants. *Journal of Child Language, 49*(4), 714-740. https://doi.org/10.1017/S0305000921000118

Kucker, S. C., McMurray, B., & Samuelson, L. K. (2015). Slowing down fast mapping: Redefining the dynamics of word learning. *Child Development Perspectives, 9*(2), 74–78. https://doi.org/10.1111/cdep.12110

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

Lindsay, D. S. (2020). Seven steps toward transparency and replicability in psychological science. *Canadian Psychology/Psychologie Canadienne, 61*(4), 310–317. https://doi.org/10.1037/cap0000222

Luk, G. (2015). Who are the bilinguals (and monolinguals)? *Bilingualism: Language and Cognition, 18*(1), 35–36. https://doi.org/10.1017/S1366728914000625

Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology, 25*(5), 605–621. https://doi.org/10.1080/20445911.2013.795574

Mattock, K., Polka, L., Rvachew, S., & Krehm, M. (2010). The first steps in word

learning are easier when the shoes fit: Comparing monolingual and bilingual infants: Phonetic variability and word learning. *Developmental Science, 13*(1), 229–243. https://doi.org/10.1111/j.1467-7687.2009.00891.x

May, L., & Werker, J. F. (2014). Can a click be a word?: Infants' learning of non-native words. *Infancy, 19*(3), 281-300. https://doi.org/10.1111/infa.12048

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review, 119*(4), 831–877. https://psycnet.apa.org/doi/10.1037/a0029872

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science, 3*(1), 24–52. https://doi.org/10.1177/2515245919900809

Morini, G., & Newman, R. S. (2019). Dónde está la ball? Examining the effect of code switching on bilingual children's word recognition. *Journal of Child Language, 46*(6), 1238–1248. https://doi.org/10.1017/S0305000919000400

Oakes, L. M. (2017). Sample Size, statistical power, and false conclusions in infant looking-time research. *Infancy, 22*(4), 436–469. https://doi.org/10.1111/infa.12186

Place, S., & Hoff, E. (2011). Properties of dual language exposure that influence 2-year-olds' bilingual proficiency. *Child Development, 82*(6), 1834–1849. https://doi.org/10.1111/j.1467-8624.2011.01660.x

R Core Team (2020). *R:* A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramon-Casas, M., Swingley, D., Sebastián-Gallés, N., & Bosch, L. (2009). Vowel categorization during word recognition in bilingual toddlers. *Cognitive Psychology, 59*(1), 96-121. https://doi.org/10.1016/j.cogpsych.2009.02.002

Rocha-Hidalgo, J., & Barr, R. (2022). Defining bilingualism in infancy and toddlerhood: A scoping review. *International Journal of Bilingualism*. https://doi.org/10.1177/13670069211069067

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641.

Schafer, G., & Plunkett, K. (1998). Rapid word learning by fifteen-month-olds under tightly controlled conditions. *Child Development, 69*(2), 309–320. https://doi.org/10.1111/j.1467-8624.1998.tb06190.x

Schott, E., Kremin, L. V., & Byers-Heinlein, K. (2022). The youngest bilingual Canadians: Insights from the 2016 census regarding children aged 0–9 years. *Canadian Public Policy, 48*(2), 254-266. https://doi.org/10.3138/cpp.2021-064

Schott, E., Rhemtulla, M., & Byers-Heinlein, K. (2019). Should I test more babies? Solutions for transparent data peeking. *Infant Behavior and Development, 54*, 166–176. https://doi.org/10.1016/j.infbeh.2018.09.010

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences, 108*(15), 6038–6043. https://doi.org/10.1073/pnas.1017617108

Sidhu, D.M., & Pexman, P.M. (2018). Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review, 25*, 1619–1643. https://doi.org/10.3758/s13423-017-1361-1

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Singh, L., Fu, C. S. L., Tay, Z. W., & Golinkoff, R. M. (2018). Novel word learning in bilingual and monolingual infants: Evidence for a bilingual advantage. *Child Development, 89*(3), e183–e198. https://doi.org/10.1111/cdev.12747

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388*(6640), 381–382. https://doi.org/10.1038/41102

Tan, S. H., & Schafer, G. (2005). Toddlers' novel word learning: Effects of phonological representation, vocabulary size and parents' ostensive behaviour. *First Language, 25*(2), 131–155. https://doi.org/10.1177/0142723705050338

Taxitari, L., Twomey, K. E., Westermann, G., & Mani, N. (2020). The limits of infants' early word learning. *Language Learning and Development, 16*(1), 1-21.

https://doi.org/10.1080/15475441.2019.1670184

Tsui, A. S. M., Byers-Heinlein, K., & Fennell, C. T. (2019). Associative word learning in infancy: A meta-analysis of the switch task. *Developmental Psychology, 55*(5), 934–950. https://doi.org/10.1037/dev0000699

Tsuji, S., Cristia, A., Frank, M. C., & Bergmann, C. (2020a). Addressing publication bias in meta-analysis: Empirical findings from community-augmented meta-analyses of infant language development. *Zeitschrift Für Psychologie*, *228*(1), 50–61. https://doi.org/10.1027/2151-2604/a000393

Tsuji, S., Jincho, N., Mazuka, R., & Cristia, A. (2020b). Communicative cues in the absence of a human interaction partner enhance 12-month-old infants' word learning. *Journal of Experimental Child Psychology*, 191, 104740. https://doi.org/10.1016/j.jecp.2019.104740

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology, 34*(6), 1289–1309. https://psycnet.apa.org/doi/10.1037/0012-1649.34.6.1289

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development, 1*(2), 197–234. https://doi.org/10.1080/15475441.2005.9684216

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy, 3*(1), 1-30. https://doi.org/10.1207/S15327078IN0301_1

Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18-month-olds. *Developmental Psychology, 30*(4), 553–563. https://psycnet.apa.org/doi/10.1037/0012-1649.30.4.553

Yin, J., & Csibra, G. (2015). Concept-based word learning in human infants. *Psychological Science, 26*(8), 1316-1324. https://doi.org/10.1177/0956797615588753

Yoshida, K. A., Fennell, C. T., Swingley, D., & Werker, J. F. (2009). Fourteen-month-old infants learn similar-sounding words. *Developmental Science, 12*(3), 412–418. https://doi.org/10.1111/j.1467-7687.2008.00789.x.

Yu, C., & Smith, L. B. (2011). What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science, 14*(2), 165-180. https://doi.org/10.1111/j.1467-7687.2010.00958.x

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. Cognition, *125*(2), 244-262. https://doi.org/10.1016/j.cognition.2012.06.016

Data, code and materials availability statement

All data, code and stimuli used in the present study are available at the Open Science Framework. Stimuli are available at https://osf.io/g6nrv. Data and scripts are available at https://osf.io/upy7f.

Ethics statement

The current study was conducted according to the Declaration of Helsinki and ethics approval was obtained from the Human Research Ethics Board of Concordia University (certification numbers UH2011-041 and 10000439). All participants' parents gave informed written consent before taking part in the study.

Authorship and Contributorship Statement

AMGB: Writing - Original Draft, Writing - Review & Editing, Formal Analysis. RDB: Writing - Review & Editing, Formal Analysis. HK: Writing - Review & Editing, Formal Analysis. KBH: Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

We are grateful to all the families who participated in this research. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada to KBH (402470-2011 and 2018-04390), a FQRSC postdoctoral fellowship to AMGB (2018-B3-205717), and a Concordia Horizon postdoctoral fellowship to RDB.

License