The development of linguistic prediction:

Predictions of sound and meaning in 2-to-5 year olds.

Chiara Gambi

Edinburgh University and Cardiff University

Fiona Gorrie

Martin J. Pickering

Hugh Rabagliati

Edinburgh University

Address for correspondence:

Chiara Gambi

School of Psychology

70, Park Place

Cardiff University

CF10 3AT Cardiff, U.K.

GambiC@cardiff.ac.uk

Phone: +44(0)29 206 88950

Abstract

Language processing in adults is facilitated by an expert ability to generate detailed predictions about upcoming words. This may seem like an acquired skill, but some models of language acquisition assume that the ability to predict is a pre-requisite for learning. This raises a question: Do children learn to predict, or do they predict to learn? We tested whether children, like adults, can generate expectations about not just the meanings of upcoming words but, also, their sounds, which would be critical for using prediction to learn about language. In two looking-while-listening experiments, we show that two-year-olds can generate expectations about meaning based on a determiner (*Can you see one…ball/two…ice-creams?*), but that even children as old as five do not show an adult-like ability to predict the phonology of upcoming words based on a determiner (*Can you see a…ball/an…ice-cream?*). Our results therefore suggest that the ability to generate detailed predictions is a late-acquired skill. We argue that prediction may not be the key mechanism driving children's learning, but that the ability to generate accurate semantic predictions may nevertheless have facilitative effects of language development.

Keywords: prediction; sound; meaning; visual-world; learning.

Word Count: 11119 (total)

10356 (main text and references, excluding tables and figures)

Data Link: https://github.com/chiara-gambi/phonpred

The development of linguistic prediction:

Predictions of sound and meaning in 2-to-5 year olds.

Introduction

A growing body of evidence suggests that we can rapidly make sense of the world thanks to prediction (Bar, 2007; Friston, 2005; Pickering & Garrod, 2013). For example, we can process sentences faster when the grammar, meaning, and sounds of upcoming words are predictable (Huettig, 2015). But prediction may do more than facilitate our ability to process the world: It may also drive learning (Dell & Chang, 2014; Elman, 1990; Rabagliati, Gambi, & Pickering, 2016). Children might learn about language, for example, by comparing their naïve expectations about upcoming words to the input, and updating their linguistic knowledge when those expectations are incorrect (i.e., in order to minimize the resulting prediction error signal). Prediction, therefore, could serve to unify processing and learning.

According to one of the most influential formulations of these ideas – predictive coding (Friston, 2005; 2010) – the mind is constantly engaged in an attempt "to match incoming sensory inputs with top-down expectations or predictions" (Clark, 2013; p.1). Detailed sensory expectations are key to this process of prediction error minimization, and hence to learning, because they allow abstract predictions to be "grounded" in a format suitable for comparison with sensory input. But while there is good evidence that detailed predictions are part of the way adults process language, it is unknown whether children generate them, and from what age. The aim of this paper is to fill this gap: We test whether children can generate predictions that are sufficiently detailed at the level of sounds that they give rise to informative error signals that may drive learning. This of course does not amount to testing whether such predictions do in fact drive learning in children, as expected under

learning-via-prediction models, but we focus on testing an important precondition or assumption of such models – namely, that children can generate detailed sound predictions.

Detailed expectations about the forms of upcoming words – i.e., more detailed than just their semantic meaning or syntactic category – can play a number of fundamental roles in learning. For example, they could help children learn about relations between linguistic structure and linguistic form, including learning about irregularities. It has frequently been suggested that children might unlearn overregularizations (e.g., the plural of *mouse* is *mice* not *mouses*) by predicting to hear one form (/maʊzɪz/) and, when they hear another form (/maɪs/), updating their internal representations (a suggestion known as implicit negative evidence; e.g., Ramscar, Dye, & McCauley, 2013; MacWhinney, 2004). This example also helps illustrate how children may generate detailed predictions even as their knowledge is still incomplete (e.g., they know how to form regular plurals but not irregular plurals). Being able to specifically predict particular words and wordforms could also be crucial for distributional learning; for example, a child who knows that *the robber* is more predictable than *the policeman* after *He will arrest* has learned something about the meaning and syntax of the verb *arrest* from its distributional properties (Gambi, Pickering, & Rabagliati, 2016; Elman, 1990). In this example, the child may leverage their emerging world knowledge (i.e., about the typical participants of arresting events) to generate detailed lexical predictions that help them learn about thematic and syntactic structure.

But while it is clear how low-level expectations about sound could drive language learning, there are reasons to believe that such detailed expectations may be too complex for young children to use. For instance, in order to be able to predict /maʊzɪz/, a child would not only need to possess a robust knowledge of the sound system, lexicon, and grammar of their native language, but would also need to be able to quickly pass information between these different levels of representation in a top-down fashion. This second point is particularly

important, because a number of studies suggest that top-down processing may be slower or more limited in young children (e.g., Snedeker & Trueswell, 2004; Snedeker & Yuan, 2008), which would cause difficulty generating detailed predictions. This, in fact, would be compatible with children's late unlearning of overregularizations (Marcus, 1995), and suggests that children may learn to predict, rather than predict to learn.

While children's ability to generate detailed predictions is unclear, there is good evidence that adults not only predict the meanings of upcoming words (e.g., Altmann & Kamide, 1999), but also their forms, including acoustic and orthographic properties (Dikker, Rabagliati, Farmer, & Pylkkänen, 2010; Herrmann, Maess, Hasting, & Friederici, 2009; Dikker & Pylkkänen, 2013; Farmer, Brown, & Tanenhaus, 2013; see also DeLong, Urbach, & Kutas, 2005, though see below). In an MEG study of reading, Dikker et al. (2010) had participants read contexts that predicted a syntactic category (e.g., noun as in *The tasteful…* or verb participle as in *The tastefully…*), followed by a target noun with visual characteristics that were typical of the orthography of nouns (e.g., *soda*) or neutral between nouns and verbs (e.g., *infant*). An enhanced sensory mismatch response occurred in visual cortex after only 100-130ms when the target's characteristics mismatched the expected syntactic category, but not when the target was neutral. Such findings suggest that adults use high-level context (e.g., the syntactic environment of a word) to derive detailed expectations about lower-level properties (see Pickering & Garrod, 2013, p. 343).

Adults' predictions are therefore detailed enough to support predictive learning (see also Gagnepin, Henson, & Davis, 2012). But what about young, language-learning children? Recent eye-tracking studies show that children do make linguistic predictions, but these studies do not specify exactly *what* children predict. For example, when two-year-olds hear a semantically-constraining verb (e.g., *eat*), they gaze towards semantically predictable pictures (e.g., of a cake, Mani & Huettig, 2012; Borovsky, Elman, & Fernald, 2012; see also Gambi et

al., 2016; Lukyanenko & Fisher, 2016). Thus, children use context to predict the message the speaker is conveying. But it is unclear if their predictive gaze is driven solely by predictions about meaning (something edible) or also by more detailed predictions about particular words (*cake*) and their component sounds.

Some evidence concerning the specificity of children's linguistic predictions comes from mismatch responses in ERP paradigms. For instance, Snedeker (2013) suggested that the mismatch response elicited when three-year-olds hear syntactic category violations (e.g., **Elle la fraise*, "she it strawberry" vs. *Elle veut manger la fraise*, "she wants to eat the strawberry"; Bernal, Dehaene-Lambertz, Millotte, & Christophe, 2010; see also Oberecker, Friedrich, & Friederici, 2005) may indicate a mismatch with low-level predictions, following Dikker et al. (2010). However, that conclusion is somewhat inconsistent with the late timing of the response (typically starting around 450ms after the onset of the violation), which instead suggests it could reflect integration difficulties. Under the latter interpretation, children may find it more difficult to process infrequent or ungrammatical word sequences, rather than actively generating expectations about upcoming word categories and their form features.

Somewhat stronger evidence that detailed low-level predictions are generated from early in life comes from the finding that infants as young as 3 months produce a MisMatch Negativity (MMN) response to auditory oddball stimuli (e.g., the last sound in the sequence /a/-/a/-/a/-/a/-/i/; Dehaene-Lambertz & Dehaene, 1994; Friederici, Friedrich, & Weber, 2002). However, in most of these studies, expectations about form are generated via low-level repetitions, and not via higher-level (e.g., syntactic) knowledge. An exception to this is the finding that infants generate a mismatch response to stimuli that are identical (/a/-/a/-/a/-/a/-/a/) when they are first familiarized to non-identical stimuli (/a/-/a/-/a/-/a/-/i/), which suggests that infants use global statistics to predict sounds (Basirat, Dehaene, & Dehaene-Lambertz,

2014). However, this mismatch response occurred later (900ms) than the standard infant MMN (typically observed between 270-370ms), and thus it may reflect higher-level processing, rather than sensory processing.

Finally, Ylinen, Bosseler, Junttila, and Huotilainen (2016) argued that infants' processing of words versus non-words provides evidence for prediction errors. They showed that Finnish 12- and 24-month-olds produced different ERP responses to two equally infrequent oddballs, the disyllables /kuk:a/ and /kuk:e/, with the former eliciting an early negativity response. /kuk:a/ is a familiar word (meaning *rooster*), but /kuk:e/ is a nonword, and Ylinen et al. argued that the distinct responses could only be explained if infants had used the first syllable to predict the second syllable, resulting in different processing depending on whether the continuation was a possible word. But this claim is hard to evaluate, partly because of one unexpected result – the oddball unfamiliar words generated an unusual mismatch *positivity* – and partly because these data could also be explained non-predictively, with infants processing familiar versus unfamiliar multi-syllable words in different ways.

While all of these studies have used mismatch responses to infer that children generate detailed form predictions, such evidence is at best quite indirect, and leaves these studies open to alternative "integration cost" explanations: Rather than actively generating expectations about sounds, children may find it more difficult to process infrequent sounds. A direct test of children's ability to predict sounds is therefore needed. If children can indeed predict sounds from early on, then it would mean that one fundamental assumption of learning-via-prediction models is correct. But if this highly sophisticated ability is slow to develop, then prediction may instead be the product of learning, rather than the driving force behind it.

Here, we provide such a test, using visual-world eye-tracking, the same paradigm that shows semantic prediction in children (e.g., Mani & Huettig, 2012). Although eye-tracking

cannot measure sound predictions directly, we designed our experiments so that, if participants generated an appropriate sound prediction, then they could predictively resolve the referent of an upcoming word (i.e., gaze to a predictable picture). In Experiment 1, children between two and five, as well as adults, viewed pairs of pictures (e.g., of a ball and an ice-cream) and heard instructions such as *Can you see a…ball/an…ice-cream?* (Figure 1A). Our displays always paired a picture whose name begins with a vowel (vowel-initial picture) with one whose name begins with a consonant (consonant-initial picture). If listeners predict sound, then looks to the vowel-initial picture should be more frequent, and increase more rapidly over time, following *an* (vs. *a*), since *an* is more usually followed by a vowel than *a*, in most varieties of English[1] (Raymond, Fisher, & Healy, 2002). Crucially, our sentences included a pause before the critical word, which gave children enough time to predict which object the speaker would refer to, before hearing the spoken name.

We chose to test the *a/an* alternation first and foremost because it is a purely phonological alternation, so if participants can use it to predict an upcoming referent, then they must be doing so based on sound rather than meaning or grammar. Our study also followed DeLong et al. (2005), who used this alternation to investigate form prediction in adults. In their experiment, adults read contexts such as *The day was breezy so the boy went outside to fly…*, followed by either *a* or *an*. The amplitude of the N400 on the determiner was larger the less expected the determiner, suggesting that adults predicted the word *kite* and its form (i.e., that it begins with a consonant), and were thus more surprised to read *an* than *a*. Converging evidence is provided by Martin et al. (2013), who found a larger N400 to unexpected than expected determiners. However, very recent evidence suggests that the N400

[1] Including the variety our participants were exposed to at the time of testing, namely Scottish English.

effect on the determiner may not be reliable (Ito, Martin, & Nieuwland, 2016; Nieuwland et al. 2017, though see Yan, Kuperberg, & Jaeger, 2017). Given the uncertainty surrounding that important finding, our study can thus also provide important evidence about form predictions in adults.

In addition, the *a/an* alternation provides a close analogue to prior child studies. For example, Lew-Williams and Fernald (2007) employed a similar method to show that 2- and 3-year-olds could use the syntactic gender of a determiner to facilitate processing of a subsequent noun: Spanish-learning children recognized *la pelota* (the ball) more quickly when the ball was the only potential referent with feminine gender (Lew-Williams & Fernald, 2007, and see Johnson, 2005 for Dutch, Van Heugten & Shi, 2009, for French). Moreover, Mahr, McMillan, Saffran, Weismer, and Edwards (2015) showed that, when English-learning 18- to 24-month-olds heard *the ball*, they gazed to the ball more quickly if *the* had been co-articulated with the onset of *ball,* compared to when there was no co-articulation. This could suggest that the co-articulatory information in *the* was used to predict the form of the subsequent noun, and thus its referent, but that study did not directly measure predictions (and children oriented to the target picture only after they had heard the target noun, *ball*); thus, co-articulation may have simply facilitated children's recognition of *ball*, rather than caused children to predict it.

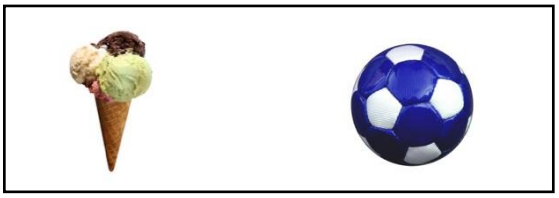
In sum, even children under two are sensitive to the information carried by determiners, including subtle co-articulatory cues, which suggests that, if they can generate sound predictions at all, then they should be able to do so based on the *a/an* alternation, which is simple (*an* precedes vowels), frequent, and relies on a salient phonological distinction (i.e., between vowels and consonants). However, there is also some evidence that children's ability to use this alternation in an adult-like way is somewhat delayed: Whereas three-year-olds never use *an* before consonants (suggesting that they appreciate the

distinction between the determiners), children will sometimes use *a* before vowels until the school years (McKee & McDaniel, 2009). This delay need not reflect incomplete knowledge of the alternation, but it might imply that only older children will predict sound in our task.

As a control condition, Experiment 1 thus compared sound-prediction to a matched meaning-prediction task. The same participants who heard sentences such as *Can you see a …ball/an… ice-cream?* also responded to sentences such as *Can you see one…ball/two…ice-creams?*, while looking at pictures of one ball (one-object picture) and two ice-creams (two-object picture; Figure 1B). If listeners use determiners (i.e., the numerals *one* and *two*) to predict numerosity, then looks to the two-object picture should be more frequent and increase more rapidly over time following *two* than *one*. We know that even two-year-olds understand *one* as exact (i.e., as meaning "one and no more than one"; Barner, Chow, & Yang, 2009), and we therefore expected all four age groups to predict numerosity. In addition, we measured children's comprehension vocabulary, and hypothesized that children with larger vocabularies would show larger prediction effects, as in previous work (e.g., Borovsky et al., 2012; Mani & Huettig, 2012).

Figure 1. Experiment 1; Sample sound-prediction (A) and meaning-prediction (B) trials.

(A) Oh! look! Can you see a… ball / an… ice-cream?



(B) Oh! look! Can you see one… ball / two… ice-creams?

Experiment 1

Method

Participants. Participants were twenty-four adult native speakers of English (M_{age}: 21.0 years, [18,24], 5 males), 40 English-learning two-year-olds (M_{age}: 29 months, [20,35], 21 males), 47 three-year-olds (M_{age}: 41 months, [36,47], 22 males), and 40 four-to-five-year-olds (M_{age}: 53 months, [48,65], 21 males). Three more children were tested but discarded because they did not complete the task, were diagnosed with language delay, or because of equipment failure. We recruited most children (121) from private nursery schools around Edinburgh, and the rest from a database of interested families. Testing continued until there were at least 40 children in each group; the larger sample size for three-year-olds is due to this group being the largest in nurseries. Ethnicity and SES were not recorded, but were representative of the area (almost entirely white, predominantly from middle-class Scottish families).

Materials and procedure. Stimuli were ten pairs of vowel-initial and consonant-initial words (*aeroplane-car*, *ant-duck*, *apple-balloon*, *arm-train*, *ear-hat*, *egg-spoon*, *elephant-dog*, *eye-bubble*, *orange-tree*, *ice cream-ball*), and corresponding pictures (see

Figure 1). Each pair was used on 4 different trials (once per determiner: *a/an/one/two*), for a total of 40. On sound-prediction trials, participants saw one exemplar of each picture. On meaning-prediction trials, they saw a one-object picture and a two-object picture. On half the trials the two-object picture was vowel initial, and on half the trials it was consonant initial (counterbalanced across lists). Position of the vowel-initial picture (left/right) was counterbalanced between items.

Each trial began with a two-second silent preview of the pictures, followed by a sentence of the form *Oh! Look! Can you see a…ball/an…ice-cream/one…ball/two…ice-creams?*, pre-recorded by a female native speaker of Scottish English with child-directed prosody. Determiners were long (*a*: 350ms; *one*, *two*, *an*: 700ms) and carried stress. Carrier phrases (e.g., *Can you see a…*) were recorded separately, and nouns were spliced into the recordings later to avoid informative co-articulatory cues on the determiner. A pause inserted before the noun ensured children had time to make predictive eye-movements. Target nouns always began approximately 1200ms after determiner onset (so *a* was followed by a longer pause than the other determiners).

Participants were instructed simply to look and listen. To engage children, the task was presented as a game in which they collected stars to obtain stickers. Two seconds after the onset of the target noun (and regardless of the child's performance), a star appeared next to the target picture, after which the experimenter placed stickers onto a chart that the child could keep at the end. Presentation order was individually randomized. Stimuli appeared on a laptop fitted with a REDn Scientific eye-tracker (SensoMotoric Instruments GmbH, www.smivision.com). The tracker was calibrated once at the start of the session and once after 20 trials using a 5-point grid. It recorded fixations binocularly at 30Hz, but we only analyzed right-eye fixations. After the listening task, children completed the British Picture

Vocabulary Scale (Second Edition, Dunn, Dunn, Whetton, & Burley, 1997), as a measure of comprehension vocabulary. Sessions lasted 20-30 minutes.

Analysis. Since sound- and meaning-prediction trials were not part of a crossed experimental design (e.g., there is no trial where *an* is used to refer to two objects), we first analyzed them separately. To test whether participants predicted sound, we modeled the proportion of looks to the vowel-initial picture; to test prediction of meaning, we modeled looks to the two-object picture.[2] Follow-up analyses compared the models to assess whether participants predicted sound to the same extent as meaning. All analyses also compared each age group (2-year-olds, 3-year-olds, 4-to-5-year-olds, and adults) to the immediately younger one using backward difference coding. Follow-up analyses examined each age group separately.

Using BeGaze (Version 3.6), we grouped raw data into fixations to areas of interest (corresponding to the right-hand side and the left-hand side of the screen), in 100ms increments, during a 1300ms window beginning at determiner onset and ending 100ms after the earliest (across items) noun onset (the 100ms delay accounts for delays in launching eye-movements; Trueswell, 2008). We discarded any trials on which no gaze was recorded (due to looking away, or track loss) on 60% or more of the samples (2-year-olds: 20.2% of trials, 3-year-olds: 18.0%, 4-5-year-olds: 9.7%, adults: 1.0%), and a further 6 trials for experimenter error.

We analyzed how participants' predictive fixations changed over time, using growth curve analyses (Mirman, 2014). There were two sets of analyses, one averaging over items

[2] Qualitatively similar results held when modelling looks to the consonant-initial picture and the one-object picture, respectively.

(i.e. with participants as random effects) and one averaging over participants (i.e., with items as random effects). The purpose of averaging is to obtain more robust estimates, and the purpose of the two sets of analyses is to be able to generalize both over participants and over items. We applied the empirical logit transformation to our averaged proportion data, as recommended by Barr (2008). Models had the structure: 1 + Determiner + Time + AgeGroup + a linear Time term, plus their full set of interactions, and a maximal random effects structure (Barr, Levy, Scheepers, & Tily, 2013), except when otherwise specified; random correlations were always set to zero to aid convergence (Bates, Kliegl, Vasishth, & Baayen, 2015).[3] The effect of Determiner captures overall differences between conditions: A positive estimate indicates participants were more likely to look at the vowel-initial (two-object) picture after *an* (*two*) than after *a* (*one*). The interaction between Determiner and the linear time term captures how gaze changed over time: A positive estimate indicates that participants looked increasingly more to the vowel-initial (two-object) picture after *an* (*two*) than *a* (*one*). Fixed effects were contrast coded and centered. Follow-up analyses of children's eye-movement data additionally investigated whether prediction varied with the size of their comprehension vocabulary (raw BPVS score), while controlling for age in months (both centered). Vocabulary scores ranged from 9 to 85, and strongly correlated with age ($r(125)=0.77$, $p<.001$).

Results

Meaning-prediction trials. We found that both children and adults could use the meanings of numbers to predictively resolve upcoming referents. As shown in Figure 2, left panel, the average proportion of looks to the two-object picture, computed over the entire prediction window, was higher after the word *two* than after *one* for adults, four-to-five-year-

[3] Models with quadratic time terms led to the same conclusions.

olds, and three-year-olds. This was confirmed by significant effects of Determiner in by-participant and by-item analyses (analysis type indicated by [p] and [i], respectively, Table 1). Only two-year-olds did not show this effect (see Table 1).[4]

However, as Figure 2 (right panel) shows, the lack of an overall effect in two-year-olds was qualified by an interaction of Determiner with Time: After hearing *two*, two-year-olds increasingly looked to the two-object picture as the trial progressed, especially towards the end of the prediction window (and the same was true for all other age groups, Table 1).[5]

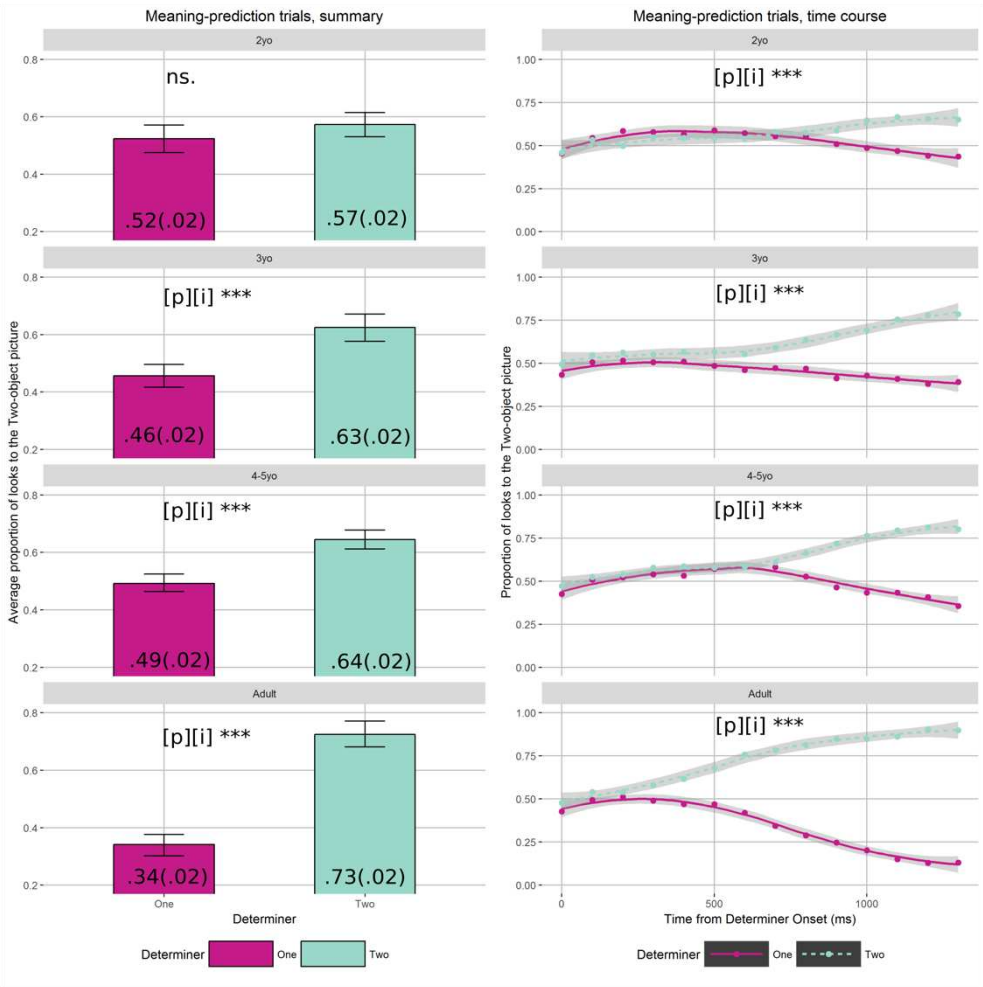
[4] The effect of Determiner did not vary with the children's vocabulary (all *p*'s >.05). By-item analyses did not include interactions between age and vocabulary in the random structure.

[5] The increase in predictive looks over time was faster for children with larger vocabularies among three-year-olds (Determiner:Time:Vocabulary, [p] Beta= .09, SE=.04, CI=[.02,.16], t=2.46; $\chi2(1)$=6.37, *p*=.01; [i] Beta= .04, SE=.01, CI=[.02,.06], t =4.22; $\chi2(1)$=11.70, *p*<.001) and four-to-five-year-olds (Determiner:Time:Vocabulary, [p] Beta= .06, SE=.03, CI=[.008,.110], t =2.26; $\chi2(1)$=5.49, *p*=.02; [i] Beta= .02, SE=.01, CI=[.005,.043], t =2.50; $\chi2(1)$=6.02, *p*=.01); there were no effects of vocabulary in 2-year-olds (all *p*'s >.05).

Table 1. Summary of critical growth-curve model effects for meaning-prediction trials, in by-participant [p] and by-item [i] analyses. From left to right: estimates (SE), 95% Confidence Intervals or CI (from the *confint* function, method="Wald"), *t* values, and chi-square and *p* values from log-likelihood ratio tests.

Age Group	Analysis	Determiner	Determiner:Time
2 yo	[p]	.16(.12), *CI*=[-.07,.39], *t*= 1.37; $\chi2(1)$=2.02, *p*=.155	.97(.29), *CI*=[.39,1.54], *t*= 3.28; $\chi2(1)$=10.52, *p*=.001
	[i]	.08(.08), *CI*=[-.08,.23], *t*= 0.99; $\chi2(1)$=1.06, *p*>.250	.62(.17), *CI*=[.28,.96], *t*= 3.59; $\chi2(1)$=11.08, *p*=.001
3 yo	[p]	.72(.14), *CI*=[.45,.99], *t*= 5.16; $\chi2(1)$=25.32, *p*<.001	1.84(.33), *CI*=[1.19,2.49], *t*= 5.53; $\chi2(1)$=28.57, *p*<.001
	[i]	.42(.06), *CI*=[.30,.54], *t*= 6.76; $\chi2(1)$=25.54, *p*<.001	1.13(.17), *CI*=[.80,1.46], *t*= 6.70; $\chi2(1)$=25.64, *p*<.001
4-5 yo	[p]	.60(.10), *CI*=[.40,.79], *t*= 6.05; $\chi2(1)$=32.82, *p*<.001	2.02(.32), *CI*=[1.39,2.64], *t*= 6.28; $\chi2(1)$=34.94, *p*<.001
	[i]	.36(.04), *CI*=[.28,.45], *t*= 8.66; $\chi2(1)$=26.83, *p*<.001	1.23(.13), *CI*=[.98,1.49], *t*= 9.42; $\chi2(1)$=31.49, *p*<.001
Adults	[p]	1.69(.14), *CI*=[1.42,1.96], *t*= 12.14; $\chi2(1)$= 68.95 , *p*<.001	4.48(.30), *CI*=[3.89,5.07], *t*= 14.87; $\chi2(1)$= 84.46 , *p*<.001
	[i]	1.20(.08), *CI*=[1.05,1.36], *t*= 15.23; $\chi2(1)$=52.63, *p*<.001	3.22(.18), *CI*=[2.88,3.57], *t*= 18.18; $\chi2(1)$=54.73, *p*<.001

Figure 2. Meaning-prediction trials. (Left panel): average proportion of looks to the two-object picture during the prediction window, after *one* (left bar) or *two* (right bar); error bars represent 95% CI computed over 1000 bootstrapped samples; superimposed numbers are means (and standard deviations). (Right panel): proportion of looks to the two-object picture over time during the same prediction window. Points are observed values; lines (solid for *one*, dashed for *two*) were obtained by non-parametric smoothing (method="loess", with 95% CI shaded in grey). *p* values correspond to the main effect of Determiner in the left panels and to the Determiner:Time interaction in the right panels (see Table 1; *** = p >.001, ** = p < .01, * = p <.05; [p] = by-participant analysis, [i] = by-item analysis).



Cross-age comparisons showed that the effect of Determiner was larger in three- than two-year-olds ([p] Beta = .50, SE = .16, CI =[.19,.82], t= 3.18, $\chi2(1)$=10.19, *p*=.001; [i] Beta=.29, SE = .10, CI =[.10,.49], t= 2.92, $\chi2(1)$=7.83, *p*=.005), and in adults than in four-to-five-year-olds ([p] Beta = 1.07, SE = .19, CI =[.70,1.44], t= 5.68, $\chi2(1)$=31.40, *p*<.001; [i] Beta = .85, SE = .08, CI =[.70,1.00], t= 10.99, $\chi2(1)$=40.03, *p*<.001), but did not differ between three- and four-to-five-year-olds (both *p*'s >.05). We also found a similar developmental pattern for the interaction of Determiner with Time, which again was larger in three- than two-year-olds (although this effect was reliable only in by-items analyses; [p] Beta = .56, SE = .42, CI =[-.26,1.38], t= 1.35, $\chi2(1)$=1.87, *p*=.171; [i] Beta = .43, SE = .12, CI =[.19,.67], t= 3.55, $\chi2(1)$=12.62, *p*<.001), and in adults than in four-to-five-year-olds ([p] Beta = 2.35, SE = .50, CI =[1.38,3.32], t= 4.73, $\chi2(1)$=22.18, *p*<.001; [i] Beta = 1.96, SE=.17, CI =[1.63,2.30], t= 11.48, $\chi2(1)$=131.68, *p*<.001), but did not differ between three- and four-to-five-year-olds (both *p*'s >.05).

Sound-prediction trials. While all age groups showed strong evidence of predicting meaning, the pattern for predictions of sound were quite different (Figure 3 and Table 2). Adults produced strong evidence of phonological predictions; they looked more to the target overall and increasingly looked to the target as the trial progressed. Two-year-olds showed no evidence of phonological predictions, either overall or over the trial[6]. Older children, by contrast, did look overall more to the vowel-initial picture after hearing *an*, as indicated by

[6] Two-year-olds with larger vocabularies did show increased looks towards the vowel-initial picture after *an* compared to after *a* (Determiner:Vocabulary: [p] Beta = .032, SE=.001, CI=[.015,.048], t =3.69; $\chi2(1)$=12.84, *p*<.001; [i] Beta = .017, SE=.006, CI=[.006,.028], t =3.07; $\chi2(1)$=8.49, *p*=.004; there were no effects of vocabulary in the other age groups, all *p*'s <.05)

significant effects of Determiner in both by-participant and by-item analyses (Table 2 and Figure 3, left panel)[7]. Comparisons across age groups showed that the effect of Determiner was larger in adults than in four-to-five-year-olds ([p] Beta= .55, SE = .16, CI =[.23,.86], t= 3.43, $\chi2(1)=11.59$ $p<.001$; [i] Beta = .43, SE = .10, CI =[.24,.63], t= 4.38, $\chi2(1)=14.61$, $p<.001$), but did not differ between children of different ages (all p's >.05).

However, although these older children showed an overall preference for the target, the temporal dynamics of their predictive gaze was not adult-like. Whereas adults looked more to the vowel-initial picture over time, just as they did on meaning-prediction trials, children did not show this effect (Figure 3, right panel). There was no significant interaction between Determiner and Time in any of the age groups (Table 2), and this did not depend on vocabulary (all p's >.05).[8] Accordingly, cross-age comparisons showed that the interaction of Determiner with Time was larger in adults than in four-to-five-year-olds ([p] Beta = 2.67, SE = .49, CI =[1.70,3.63], t= 5.42, $\chi2(1)=28.74$, $p<.001$; [i] Beta = 2.07, SE = .18, CI =[1.72,2.41], t= 11.75, $\chi2(1)=137.77$, $p<.001$), but did not differ between children of different ages (all p's >.05).

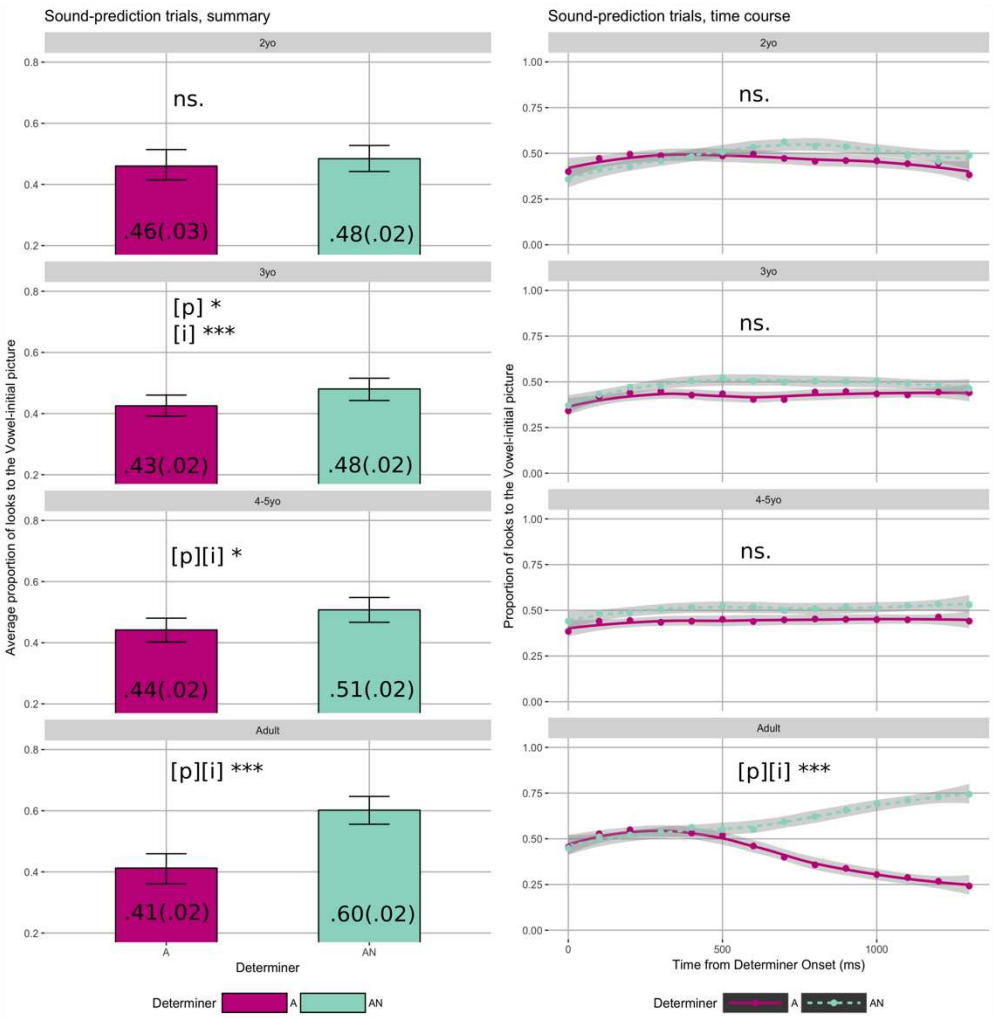
[7] By-item analyses did not include interactions between age and vocabulary in the random structure.

[8] If anything, the Determiner:Time interaction was slightly smaller in three-year-olds with larger vocabularies, although only significant in by-items analyses ([p] Beta = -.06, SE = .03, CI =[-.12,.01], t=-1.69, $\chi2(1)=3.04$, p=.081; [i] Beta =-.04, SE =.02 , CI =[-.08,-.01], t=-2.25 $\chi2(1)=4.75$, p=.029).

Table 2. Summary of critical growth-curve model effects for sound-prediction trials.

Age Group	Analysis	Determiner	Determiner:Time
2 yo	[p]	-.02(.09), *CI*=[-.20,.16], *t*= -0.23; $\chi2(1)$=0.06, *p*>.250	.31(.32), *CI*=[-.31,.93], *t*= 0.98; $\chi2(1)$=1.06, *p*>.250
	[i]	-.02(.06), *CI*=[-.14,.11], *t*= -0.26; $\chi2(1)$=0.07, *p*>.250	.10(.20), *CI*=[-.30,.49], *t*= 0.48; $\chi2(1)$=0.26, *p*>.250
3 yo	[p]	.21(.10), *CI*=[.02,.40], *t*= 2.18; $\chi2(1)$=5.08, *p*<.05	.24(.31), *CI*=[-.37,.86], *t*= 0.78; $\chi2(1)$=0.66, *p*>.250
	[i]	.15(.04), *CI*=[.08,.22], *t*= 4.26; $\chi2(1)$=11.49, *p*<.001	.04(.13), *CI*=[-.23,.30], *t*= 0.28; $\chi2(1)$=0.08, *p*>.250
4-5 yo	[p]	.21(.09), *CI*=[.03,.38], *t*= 2.33; $\chi2(1)$=5.50, *p*=.02	.07(.35), *CI*=[-.62,.76], *t*= 0.20; $\chi2(1)$=0.04, *p*>.250
	[i]	.14(.06), *CI*=[.03,.25], *t*= 2.48; $\chi2(1)$=5.19, *p*=.02	.01(.14), *CI*=[-.27,.29], *t*= 0.07; $\chi2(1)$=0.003, *p*>.250
Adults	[p]	.75(.14), *CI*=[.48,1.02], *t*= 5.45; $\chi2(1)$=23.89, *p*<.001	2.80(.32), *CI*=[2.17,3.43], *t*= 8.68; $\chi2(1)$=46.52, *p*<.001
	[i]	.57(.09), *CI*=[.38,.75], *t*= 6.09; $\chi2(1)$=17.30, *p*<.001	2.09(.32), *CI*=[1.47,2.72], *t*= 6.57; $\chi2(1)$=24.36, *p*<.001

Figure 3. Sound-prediction trials. (Left panel): average proportion of looks to the vowel-initial picture after *a* (left bar) or *an* (right bar) during the prediction window. (Right panel): proportion of looks to the vowel-initial picture over time during the same time window (solid lines for *a*, dashed lines for *an*). *p* values are the same as in Table 2. All other details as in Figure 2.



A direct comparison between our experimental conditions (Table 3) confirmed that the rate at which looks to the target increased over time (i.e., the interaction of Determiner and Time) was smaller on sound-prediction trials than on meaning-prediction trials, in all age groups except two-year-olds. Additionally, the overall prediction effect was greater for

meaning than sound in all age groups apart from two-year-olds.[9] In sum, children and adults showed strong effects of predicting meaning, and adults showed strong evidence of predicting form. But children only showed limited evidence of predicting form, in a fashion that was qualitatively different from their prediction of meaning.

[9] By-items analyses of children data did not include interactions between age and vocabulary in the random structure. The 4-to-5-year-old model did not contain any random terms for either age or vocabulary to aid convergence, while the 2-year-old model did not include the random interactions between Trial Type:Determiner or Trial Type:Determiner:Time and either age or vocabulary.

Table 3. Summary of critical growth-curve model effects for the comparison between sound- and meaning-prediction trials.

Age Group	Analysis	Determiner:Trial Type	Determiner:Time:Trial Type
2 yo	[p]	-.18(.15), *CI*=[-.47,.11], *t*=-1.20; $\chi2(1)$=1.57, *p*=.211	-.66(.43), *CI*=[-1.51,.19], *t*=-1.52; $\chi2(1)$=2.52, *p*=.112
	[i]	-.09(.09), *CI*=[-.27,.10], *t*=-0.93; $\chi2(1)$=0.91, *p*>.250	-.51(.26), *CI*=[-1.025,-.004], *t*=-1.98; $\chi2(1)$=4.12, *p*=.042
3 yo	[p]	-.51(.17), *CI*=[-.83,-.18], *t*=-3.04; $\chi2(1)$=9.66, *p*=.002	-1.58(.39), *CI*=[-2.35,-.81], *t*=-4.00; $\chi2(1)$=14.19, *p*<.001
	[i]	-.26(.06), *CI*=[-.38,-.15], *t*=-4.42; $\chi2(1)$=12.20, *p*<.001	-1.10(.21), *CI*=[-1.51,-.68], *t*=-5.18; $\chi2(1)$=22.69, *p*<.001
4-5 yo	[p]	-.44(.14), *CI*=[-.71,-.17], *t*=-3.19; $\chi2(1)$=10.57, *p*=.001	-1.98(.47), *CI*=[-2.91,-1.05], *t*=-4.16; $\chi2(1)$=18.16, *p*<.001
	[i]	-.25(.07), *CI*=[-.39,-.11], *t*=-3.46; $\chi2(1)$=10.34, *p*<.001	-1.23(.19), *CI*=[-1.61,-.86], *t*=-6.47; $\chi2(1)$=25.92, *p*< 001
Adults	[p]	-.94(.19), *CI*=[-1.32,-.56], *t*=-4.89; $\chi2(1)$=21.48, *p*<.001	-1.68(.36), *CI*=[-2.39,-.98], *t*=-4.67; $\chi2(1)$=18.66, *p*<.001
	[i]	-.64(.10), *CI*=[-.83,-.44], *t*=-6.30; $\chi2(1)$=19.53, *p*<.001 (i)	-1.13(.33), *CI*=[-1.79,-.47], *t*=-3.38; $\chi2(1)$=10.57, *p*=.001 (i)

Discussion

Experiment 1 tested whether children can use an indefinite determiner (*a/an*) to predict the form of an upcoming word, and compared this to predictions about meaning based on numerals (*one/two*). Children were clearly able to generate predictions based on meaning. The numerals constrained predictive looks even in two-year-olds, and three-to-five year-olds produced a pattern of predictive gazes that was almost adult-like (i.e., showing both an overall difference in predictive looks between numerals and an increase in this difference over time).

In contrast, predictions based on form showed a marked developmental change. Adults could clearly use the form of the indefinite determiner to predict the upcoming word, while two-year-olds showed no evidence of predicting form at all. Children older than three, however, displayed an unexpected pattern. When they heard *an*, these children were – overall – more likely to look at the vowel-initial picture (*ice-cream*), but this tendency to gaze toward the predicted picture did not increase over the trial as it did for adults, and as it did for all ages on meaning-prediction trials. Instead children showed a small but stable preference for the phonologically predictable picture throughout the prediction window.

This result may appear to provide evidence for form-based predictions in young children. However, because the small-but-stable preference that we observed in children was so different from the steady rise we observed in adults, it deserves additional scrutiny. Importantly, this type of gaze behavior is quite distinct from the findings of most eye-tracking studies of word recognition and prediction, which tend to show an increase in looks to the upcoming referent over time, as listeners process more of the linguistic input.

One possibility is that this small prediction effect is driven by predictive processing that is less stable in children (e.g., across listeners, or across items), resulting in an overall

effect, but no large increase in accuracy over time. Alternately, the effect may be caused by a looser linkage between children's eye movements and their incremental uptake of linguistic information over time. However, it is also possible that this small-but-stable difference between conditions is instead the result of unexpected baseline differences. For example, visual inspection (see Figure S1) suggested that children tended to look more at the vowel-initial picture even before the determiner *an* was actually said. This baseline difference could reflect random variation in sampling, but could also index something more systematic. In particular, Experiment 1 used a yoked picture design with repetitions, in which each pair of pictures (e.g., ball and ice-cream) was presented twice in each condition, with the identity of the target switching between presentations. Such a design is common in the field (e.g., Lew-Williams & Fernald, 2007; Mahr et al., 2015; see Supplementary material for a fuller discussion of this), because children's vocabularies are so limited. However, it opens up the possibility that children might systematically track which pictures have been mentioned, and preferentially gaze to unmentioned pictures, causing a bias in favor of the target on its second presentation. Consistent with this, visual inspection of Figure S3 (left panel) suggested that children may not have predicted the target on those trials where they encountered a pair of pictures for the first time. By contrast, adults did do so, and both adults and children did predict the target when trials were first presented in the semantic condition (see Figure S2, left panel).

These considerations mean that, while Experiment 1 provided strong evidence that both children and adults could use meaning to generate expectations about upcoming words, and that adults could also generate form-based expectations, the evidence that 3-to-5-year-olds could generate form-based expectations was more equivocal: The pattern of findings could reflect genuine form prediction, or could indicate children's ability to track labeling episodes in our experiment. Thus, Experiment 2 used a modified design to provide an

unambiguous test of the hypothesis that children can generate form-based expectations about upcoming words.

Experiment 2

We modified Experiment 1 in two critical ways. First, we systematically manipulated whether children could predict an upcoming word based only on its form, or could instead rely on which pictures had, and had not, been previously labeled. We did this by controlling trial order such that there were equal numbers of New trials (where neither the target nor the non-target picture had been previously labeled) and Old trials (where the non-target picture had been previously labeled).

Second, we replaced meaning-prediction trials with a new set of sound-prediction trials, Same Onset trials, in which the names of both pictures (e.g., ball and tree) began with the same type of phoneme (e.g., a consonant, similar to Lew-Williams & Fernald, 2007, and Mahr et al, 2015). The logic here is that, if children can indeed use the determiner to generate form-based predictions, then they should be more likely to launch predictive looks to the target picture (ball) when they hear the determiner on Different Onset trials (ball and ice-cream) than on Same Onset trials (ball and tree), and they should also be faster to recognise the target picture name (*ball*), when it occurs on Different Onset trials than on Same Onset trials. Moreover, the contrast between children's performance on Same Onset and Different Onset trials can additionally inform us about children's ability to track an experiment's structure. If they can, then they should be able to predict the target even on Same Onset trials, where the determiner itself is uninformative.

In sum, if children indeed generate form-based expectations, then they should show more predictive looks to the target picture on Different Onset trials (ball and ice-cream) than Same Onset Trials (ball and tree), and should also be faster to recognize the noun on those

Different Onset trials. Moreover, if children can track which picture has been labeled, then they should show increased predictive looks to the target on Old trials, where they can predict which picture will be named from the structure of the experiment.

Method

Participants. Participants were 41 three-year-olds (M_{age} : 43 months, [36,47], 18 males), and 41 four-to-five-year-olds (M_{age}: 54 months, [48,69], 23 males). We did not test 2-year-olds, as Experiment 1 found no evidence that they predict sounds. An additional 7 three-year-olds and 6 four-to-five-year-olds were excluded because they did not pay attention to the task, did not finish the task, were difficult to calibrate, or because of equipment failure. We recruited 42 children from nursery schools and libraries around Edinburgh, and the remainder from a lab database. We continued testing until there were at least 40 children in each group. Ethnicity and SES were not recorded, but were representative of the area as in Experiment 1.

Materials and procedure. We used the same twenty words and pictures as in Experiment 1; again, there were two blocks of 20 trials each. Each picture appeared twice per block (once as the named target). On half the trials, the two pictures had English names beginning with the same type of phoneme (i.e., both consonant- or vowel-initial; Same Onset), so that the determiner was appropriate for both names; on the other half, the picture names began with different phonemes (one consonant- and one-vowel initial; Different Onset), so the form of the determiner was informative. Within each Onset condition, half the trials named the vowel-initial picture and half named the consonant-initial picture. Position of the target (named) picture (left/right) was counterbalanced between items.

Unlike in Experiment 1, each time a picture appeared it was paired with a different picture (i.e., we never repeated picture pairs). Picture pairings and presentation order were pseudorandomized within each block so that half of the trials were New and half were Old,

and this was the case within both the Same Onset and the Different Onset condition. On New trials, neither the target nor distractor picture had been previously labeled (within the same block), but on Old trials the distractor picture had been previously labeled (within the same block), and so children could guess the target's identity if they were tracking labeling episodes.

Within each block, we further counterbalanced which Onset condition a picture appeared in on first presentation and, within each Onset condition, whether a picture was the named target or the distractor on first presentation. Vowel-initial and consonant-initial pictures were equally likely to appear first as target or distractor. Ten pictures appeared first as target or distractor in both blocks, whereas for the other 10, the role on first presentation was switched between blocks. The order of presentation of blocks was counterbalanced across participants, as was the pseudorandomized order of presentation of trials (forward or backward). In any case, there were no more than three consecutive trials on which the target shared the same type of initial phoneme or the same condition (Different vs. Same Onset) was repeated.

Every other aspect of the procedure was the same as in Experiment 1, and we also used the same pre-recorded sentences. However, children did not complete the vocabulary test, since it had little predictive power in Experiment 1

Analysis. We only analyzed right-eye fixations, using growth curve analyses as in Experiment 1, but with differently structured models to account for the design of Experiment 2. Whereas in Experiment 1 the key comparison was between predictive looks to the vowel-initial picture as a function of the determiner (*an/a*), in Experiment 2 there were two key comparisons: 1) whether children used the structure of the experiment (New versus Old trials) to generate expectations about referents, and 2) whether children took advantage of

Different Onset trials, but not Same Onset trials, to generate predictions about form, and more quickly recognise the target word. Our models thus collapsed across determiners (*an/a*) and analyzed fixations to the target (i.e., named) picture.

We first analyzed children's predictions, using the same time window as Experiment 1. We measured the effect of New versus Old trials by fitting a model with the structure: 1 + Trial Type + a linear Time term + Trial Type:Time, plus a main effect of Age group and interactions between Age group and all of the other terms (and random effects). To analyze whether children could generate expectations on Different Onset trials but not Same Onset trials, we fit a model with the structure 1 + Onset + a linear Time term + Onset:Time, plus a main effect of Age group and interactions between Age group and all of the other terms (and random effects). We fit the latter model separately to New and Old trials, with the expectation that children need not rely on phonological predictions in Old trials, because they could guess the target from memory. In all models, fixed effects were contrast coded and centered, and random effects structure was maximal (Barr et al., 2013), with correlations set to zero (Bates et al., 2015).

We then assessed word recognition, in a window that began where the prediction window ended and lasted until 1s after the earliest noun onset. Here, we analyzed the effect of Different versus Same Onset trials, using a growth curve model that had the same structure as above.

Since there were no significant interactions with the linear time term in any of the analyses, our graphs only display average proportions across the whole Prediction (Figure 4) or Recognition window (Figure 5), but full time course graphs can be found in the Supplementary materials (Figures S4 and S5). We again discarded trials on which no gaze was recorded on at least 60% of the samples (3-year-olds, prediction window: 11.6%,

recognition window: 8.1%; 4-5-year-olds, prediction: 8.7%, recognition: 6.9%), and a further trial for experimenter error.

Finally, we conducted an additional analysis following Lew-Williams and Fernald (2007) that measured the latency of the first fixation to the target picture after determiner onset, and whether this was earlier on Different Onset than on Same Onset trials. Trials were included in this analysis only if: (1) at determiner onset, the child was not fixating the target picture, and (2) the first fixation to the target began at least 100ms after determiner onset and no later than 1100ms after noun onset. After applying these criteria, 56.4% of trials were left for analysis. We used a linear mixed-effect model with a Gaussian link function and maximal random structure for both participants and items (but with random correlations set to zero). We again analyzed new trial and old trials separately. The model's fixed effects had the following form: 1 + Onset*Age.

Results

Prediction window – Growth curve. We first analyzed if children were indeed tracking labeling episodes, and predicting which picture would be mentioned based on memory. Children were much more likely to look at the target on Old than on New trials ([p] Beta =-.46, SE =.05, CI=[-.56,-.36], t=-9.12, $\chi2(1)$ =64.82, p<.001; [i] Beta = -.48, SE =.07, CI=[-.63,-.34], t=-6.49, $\chi2(1)$ =23.42, p<.001), confirming that they could guess the target based on the previous trials.

We next asked whether children were also using the form of the determiner to predict, focusing on those trials where they could not rely on memory (i.e., New trials). Our growth curve analysis did indeed provide evidence of prediction. We replicated the small-but-stable effect of prediction from Experiment 1, such that children looked more to the target on Different Onset trials (i.e., when the determiner was predictive) than on Same Onset trials.

However, this effect was only reliable in the by-participant analysis (Beta = .13, SE =.06, CI=[.01,.25], t=2.17, $\chi2(1)=4.69$, p = .03; see Figure 4), and not by items (Beta = .12, SE = .15, CI=[-.17,.41], t=0.80, $\chi2(1)=0.66$, p>.250). [10] Also as in Experiment 1, we did not find any evidence that children's phonological predictions grew stronger over time (i.e., no interactions between Onset and Time, p > .250).

On Old trials, where children could predict based on memory, they did not show effects of form-based prediction, either overall (Onset, [p] Beta =-.01, SE =.08, CI=[-.16,.14], t=-0.16; [i] Beta = -.01, SE =.11, CI=[-.22,.19], t=-0.14), or over time (i.e., no interactions between Onset and Time, p > .250).

One important question here is why Experiment 2 showed evidence of form-based prediction only in the by-participant analysis, but not the by-items analysis. Interestingly, inspection of the random effect structure of the by-item model suggested that the by-participant effect was driven by a subset of items (8 out of 20) which showed a large effect of Onset on New trials[11]. This suggests that the prediction effect in this Experiment was driven

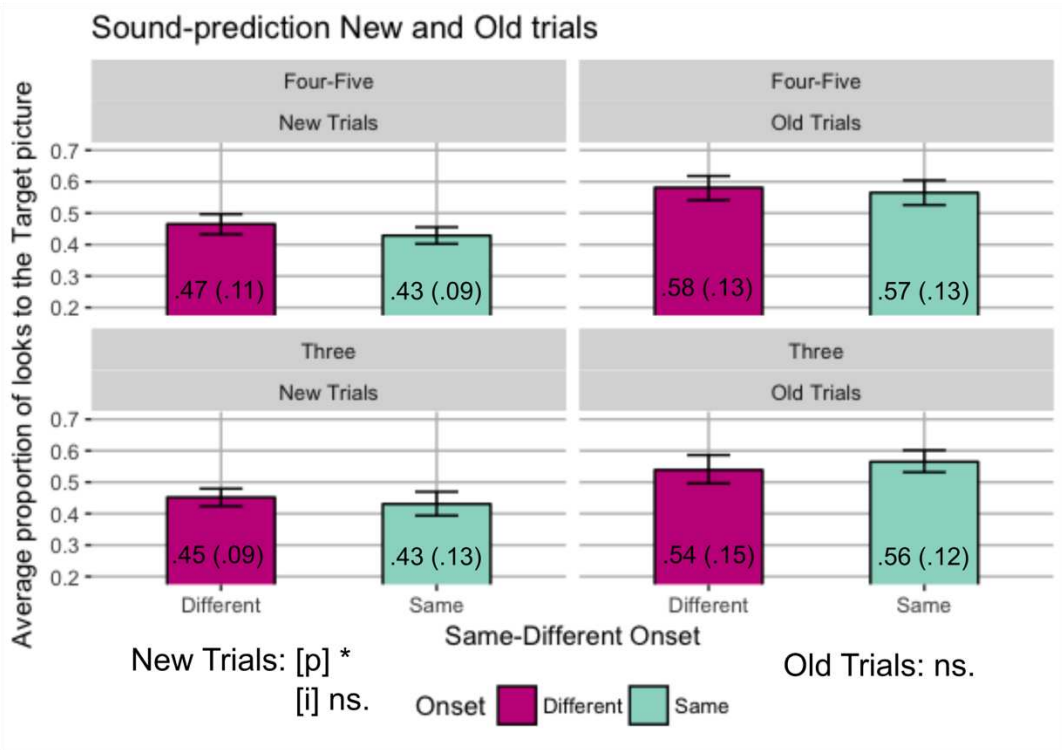
[10] Additional analyses comparing children's predictions in the first versus the second half of the experiment found no difference (Onset by Half interaction: by-participants, t = -1.60; by-items, t = -0.01); thus, children's predictions were weak throughout the experiment, and not just after encountering a number of (Same Onset) trials where prediction was not useful in identifying the target.

[11] Defined as Beta > .50 on the log-odds scale, which corresponds to 1.65 on the linear odds scale, i.e., the odds of looking at the target picture were 1.65 times higher on Different Onset trials than Same Onset trials for these items. The 8 target nouns were: *aeroplane*, *ant*, *ball*, *dog*, *egg*, *ice-cream*, *orange, train*.

by differential processing of a few items (which tended to be vowel-initial[12], 5 out of 8); we return to the implications of this point in the Discussion.

[12] It is unlikely this finding occurred because vowel-initial nouns were easier. If anything, they were more likely to be long compared to consonant-initial nouns (5/10 vs. 2/10 polysyllabic nouns). In addition, all nouns were highly frequent in a corpus of subtitles of UK TV programs for pre-schoolers (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014).

Figure 4. Experiment 2, New and Old sound-prediction trials. Prediction window. (Left panel): average proportion of looks to the target picture on New trials, when the two pictures had different onset names (left bar) or same onset names (right bar); error bars represent 95% CI computed over 1000 bootstrapped samples; superimposed numbers are means (and standard deviations). (Right panel): proportion of looks to the target picture during the same time window on Old trials. *p* values (also reported in text) indicate the effect of Onset. Plots of gaze over time can be found in the supplement.



Recognition window – Growth curve. If children use the form of the determiner to predict an upcoming word, then they should also be faster to recognize the target word. Consistent with this, on New trials, children looked more to the target on Different Onset trials than Same Onset trials. Again, however, this effect was reliable only in the by-participant analysis (Beta = .16, SE =.07, CI=[.03,.29], t=2.37, $\chi2(1)=5.67$, p = .02; Figure 5), and not the by-items analysis (Beta = -.001, SE = .016, CI=[-.030,.028], t=-0.06, $\chi2(1)=0.003$, $p>.250$). Unlike for the prediction window analyses, there was no clearly

identifiable subset of items driving the effect. There was no interaction of Onset by Time ([p], Beta =.24, SE =.16, CI=[-.06,.55], t=1.55, $\chi2(1)=2.42$, p=.120; [i] Beta =.22, SE =.28, CI=[.22,1.27], t=0.79, $\chi2(1)=0.62$, p>.250). As expected, on Old trials, there was no effect of Onset (Onset, [p] Beta = -.01, SE =.06, CI=[-.13,.11], t=-0.16; [i] Beta = .06, SE =.03, CI=[-.004,.117], t=1.82; Onset:Time, [p] Beta =.22, SE =.16, CI=[-.08, .53], t=1.43; [i] Beta =-.41, SE =.44, CI=[-1.27,.44], t=-0.94).

Recognition window - First fixation latency. Unlike in the growth curve analyses, an analysis of first fixation times did not uncover an effect of Onset on either New (B=-25 ms, SE=68, CI=[-158,108],t=-0.37) or Old trials (B=63 ms, SE=47, CI=[-29,154],t=1.34).

Figure 5. Gaze to Target during Recognition window for Experiment 2. See Figure 4 for details. Plots of gaze over time can be found in the supplement.

Discussion

In Experiment 2, 3-to-5 year olds showed some limited ability to use the form of the determiner (*a* vs. *an*) to predict an upcoming noun. We also found some indication, from growth curve analyses, that hearing a predictive determiner facilitated children's recognition of a word (i.e., enhanced looking to the target picture once the word was said), although this effect was not confirmed in first fixation latency analyses. Finally, we found good evidence that children could track the structure of the experiment to predict which picture would be the target, although this finding did not explain their apparent ability to predict using the determiner.

Importantly, and as we emphasized, the effect of predictive determiners on children's prediction and recognition of words was, at best, weak and inconsistent, with a small subset of items apparently driving the effect. Interestingly, such items were (mostly) vowel-initial, which suggests a potential explanation. Some children may have performed better with these items because they had misanalysed the picture names (e.g., treating *an ice-cream* as *a nice-cream*, and storing the latter as a lexical entry). If so, when children heard the nasal in *a nice-cream*, they may have interpreted it as the onset of the target noun, and this might in turn have directed their attention towards the target picture. In sum, these analyses suggest that, up until the age of 5, children do make some predictive use of the determiners *a* and *an*, but it seems likely that this occurs on a noun-by-noun basis rather than across the whole lexicon.

General Discussion

Two visual-world experiments investigated whether children are able to generate low-level expectations about the forms of upcoming words. Experiment 1 showed that adults can

use the form of a determiner to generate expectations about upcoming sounds, but the data for children was more equivocal; in contrast, we found strong evidence that both children (even 2-year-olds) and adults could generate high-level, semantic expectations about upcoming words. Experiment 2 confirmed that 3-to-5 year olds generate some predictions about the forms of upcoming nouns, but it also confirmed that the pattern of eye movements elicited by these predictions is quite distinct from the patterns elicited by semantic predictions (and by form predictions in adults). This suggests that there are important limitations to children's ability to generate low-level predictions which, as we discuss, have implications for theoretical models of prediction and language learning.

Is Prediction a Fundamental Component of both Language Processing and Learning?

In the Introduction, we argued that detailed predictions about low-level features such as sound are a core assumption of any account which proposes that language develops through a process of online, prediction-driven learning. This is because predictions need to be grounded in a format that allows them to be compared against the input, even when the comprehender is learning about higher-level properties such as syntax or semantics, and especially when learning about the form of words (e.g., when learning irregular plural forms in English; see, for instance, Ramscar et al., 2013).

Our data show that, even at age 5, children have limited skill at using linguistic context to predict form. Although children showed overall above-chance looking towards the target, they did not appear to predict the target consistently across all nouns, nor in a way that is tightly linked to the incremental uptake of linguistic information, as shown by their failure to increasingly look to the target over time. Why do children show such limited effects of form-based prediction, and what does this mean for predictive models of learning?

One possible explanation of the data pattern is that children's form-based expectations are very inaccurate (e.g., because they have limited knowledge of the *a/an* alternation). This explanation would preserve the idea that form-based expectations may play an important role in learning, but seems unlikely because it also implies that older and more knowledgeable children should show better performance in our task. However, we observed little in the way of developmental change between older and younger children. A second possible explanation is that children predict the sounds of upcoming words, but do not use those predictions for the task of guessing the words' referents. However, this explanation also seems unlikely, because 18-month-olds are already able to guess the referents of spoken words on the basis of partial phonological information (e.g., showing an increase in looking towards a cat vs. a dog from the very moment they hear the onset of the word *kitty*; Fernald, Swingley, & Pinto, 2001).

Thus, we suggest that a third explanation for our findings is the most likely. Children's expectations about form are not inaccurate, or inconsequential for referent selection, but rather they are typically absent during sentence processing. As children listen to a sentence unfold, they generate expectations about the meanings of upcoming words, but they do not generate expectations about the forms of those words. Given adults can do the latter, the ability to generate online expectations about form is thus a skill that develops as we become expert language users, rather than a skill that makes language learning possible.

Under this account, the limited effects of form prediction that we observed in children do not reflect true predictions, but rather misanalyses of some nouns by some children, e.g., representing *an ice-cream* as *a nice-cream*. Such misanalyses may reflect children's attempts at chunking their input (McCauley & Christiansen, 2014b), and constitute the basis on which children later develop an adult-like ability to predict the noun based on the form of the determiner. However, while these misanalyses biased some children to gaze to the correct referent as they heard the determiner, that bias appeared to dissipate during the pause before

the target noun, presumably because children received no additional linguistic information confirming their initial referent choice during that time; the result was a small preference which did not increase over time.

We would argue that this conclusion imposes important constraints on the role that prediction plays in how children learn language. Most prediction-driven models assume that language learning occurs incrementally (word-by-word; e.g., Elman, 1990): As sentences unfold, the learner generates expectations about each upcoming word, derives a prediction error signal, and then updates an internal representation of the language to minimize such error. Our data suggest that children's incremental expectations may not be detailed enough to support this type of learning.

So, how do children learn? Instead of updating their expectations after every word in a fully incremental fashion, children may rely on more offline processes, such as hypothesis testing (e.g., Perfors, Tenenbaum, & Regier, 2011) or analogy (e.g., Gentner & Namy, 2006) during early learning. Alternatively, children may learn as they process language, but they may do so by incrementally updating statistical information about the co-occurrence of lexical items (McCauley & Christiansen, 2014a; Chater, McCauley, & Christiansen, 2016). This type of learning does not require children to generate detailed predictions based on higher-level knowledge and compare them to the input, but it can account for how children gradually develop the ability to generate top-down expectations of an increasing abstract nature (Christiansen & Chater, 2016).

Importantly, it is still possible that children's partially-developed prediction skills may facilitate their language learning, though in different ways. For example, predictions might lessen the processing burden that children face when trying to simultaneously learn and comprehend their language (e.g., Fernald, Marchman, & Hurtado, 2008a; Fernald et al., 2001;

Gambi et al., 2016; Lew-Williams & Fernald, 2007). There is strong evidence that children make predictions about the semantics of words (Mani & Huettig, 2012; Borovsky et al., 2012; and Experiment 1 in this paper), and such predictions could faciliate fast and efficient comprehension, thus freeing up cognitive resources, which would allow children to focus on processing novel words and structures (Fernald, et al., 2008a) and even guessing their semantics, thus increasing the chances of encoding their meanings (e.g., Ferguson, Graf, & Waxman, 2014).

There are of course some limitations to our conclusions. For example, the *a/an* distinction is just one example of form-based prediction, and it may represent a relatively weak predictive cue (see Ito et al., 2016); it will therefore be important to investigate form-based prediction in other linguistic phenomena. Similarly, future studies may investigate at what age children begin making form predictions based on this alternation, and what other skills this may depend on (e.g., literacy). But overall, our findings do not accord with the idea that prediction drives language learning in young children; rather, we suggest prediction may facilitate language processing, which in turn may enhance the efficacy of learning.

In sum, our findings support the idea that children must learn to predict (Rabagliati et al., 2016). Although children possess the ability to generate high-level expectations from early on, the main role of this ability may be to facilitate language processing (and, only indirectly, support learning). As they learn more and more about the complex patterns in their language, children might eventually also acquire the sophisticated, expert adult skills necessary to generate low-level expectations on a word-to-word basis. But such detailed low-level expectations do not appear to drive language development in the preschool years, which runs counter to the idea that prediction is the basis of learning.

Methodological Implications: The Dangers of Item Repetitions

The results of both Experiments 1 (post-hoc analyses) and 2 (old vs. new trials) suggested that children as young as three tracked labeling episodes over the course of our experiment, and were using that information to guess which picture would be labeled next. This finding is important because item repetitions are present in many studies of children's word recognition (see Fernald, Zangl, Portillo, & Marchman, 2008b and our Supplementary Material). Fortunately, however, there are reasons to believe the effects of this confound have been minimal. First, many studies have involved children younger than 2, and we did not find evidence for tracking in 2-year-olds. In addition, many studies used multiple exemplars for each depicted word meaning (e.g., using a variety of depicted balls), and this variability may help reduce children's tracking (though this is yet to be tested). Nevertheless, we recommend that future studies, particularly those with older children, use a Latin Square design to avoid item repetition entirely, or alternatively introduce as much variability as possible to minimize tracking.

Conclusion

While young children robustly generate predictions about meaning, our study suggests that they do not use linguistic context to generate strong and consistent predictions about form, even by the age of five. We have argued that this finding is inconsistent with theories of language acquisition in which children learn from incrementally-generated prediction errors. We suggest that high-level expectations may facilitate children's learning by minimizing the difficulty of language processing, but prediction cannot be the main driver of children's language learning; rather, the ability to generate low-level expectations is a skill that must itself be learned by children.

Acknowledgements

This project was supported by a Leverhulme Trust Research Project Grant (RPG-2014-253) to Hugh Rabagliati and Martin Pickering. and an ESRC Future Research Leaders award (ES/L01064X/1) to Hugh Rabagliati. We thank all the participating children, families, and nurseries.

References

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247-264.

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences, 11*(7), 280-289.

Barner, D., Chow, K., & Yang, S.-J. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive Psychology, 58*(2), 195-219.

Barr, D. J. (2008). Analyzing 'visual world'eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59(4),* 457-474.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278.

Basirat, A., Dehaene, S., & Dehaene-Lambertz, G. (2014). A hierarchy of cortical responses to sequence violations in three-month-old infants. Cognition, 132(2), 137-150.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). Parsimonious mixed models. Retrieved from http://arxiv.org/pdf/1506.04967.pdf

BeGaze [Computer software]. (2016). SensoMotoric Instruments GmbH, www.smivision.com

Bernal, S., Dehaene-Lambertz, G., Millotte, S., & Christophe, A. (2010). Two-year-olds compute syntactic structure on-line. Developmental science, 13(1), 69-76.

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology, 112*(4), 417-436.

Chater, N., McCauley, S.M. & Christiansen, M.H. (2016). Language as skill: Intertwining comprehension and production. *Journal of Memory and Language, 89,* 244-254.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39*, e62. doi:doi:10.1017/S0140525X1500031X

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(2), 181-204.

Dehaene-Lambertz, G., & Dehaene, S. (1994). Speed and cerebral correlates of syllable discrimination in infants. *Nature, 370(6487),* 292-295.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1634), 20120394.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117-1121.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2017). Is there a replication crisis? Perhaps. Is this an example? No: a commentary on Ito, Martin, and Nieuwland (2016). *Language, Cognition and Neuroscience, 1-8.*

Dikker, S., Rabagliati, H., Farmer, T. A., & Pylkkänen, L. (2010). Early occipital sensitivity to syntactic category is based on form typicality. *Psychological Science, 21*(5), 629-634.

Dikker, S., & Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain and language, 127(1)*, 55-64.

Dunn, L., Dunn, L., Whetton, C., & Burley, J. (1997). The British Picture Vocabulary Scale—Second Edition. Windsor: NFER-Nelson.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179-211.

Farmer, T. A., Brown, M., & Tanenhaus, M. K. (2013). Prediction, explanation, and the role of generative models in language processing. *Behavioral and Brain Sciences, 36*(03), 211-212.

Ferguson, B., Graf, E., & Waxman, S.R. (2014). Infants use known verbs to learn novel nouns: Evidence from 15- and 19-month-olds. *Cognition, 131*, 139-146.

Fernald, A, Marchman, V.A., & Hurtado, N. (2008a). Input affects uptake: How early language experience influences processing efficiency and vocabulary learning. Paper presented at the 7th IEEE International Conference on Development and Learning, Monterey, CA.

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.

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008b). Looking while listening: Using eye movements to monitor spoken language. In I. A. Sekerina, E. M. Fernandez, & H. Clahsen (Eds.), *Developmental psycholinguistics: On-line methods in children's language processing* (Vol. 44, pp. 97-135). Amsterdam/Philadelphia: John Benjamins.

Friederici, A. D., Friedrich, M., & Weber, C. (2002). Neural manifestation of cognitive and precognitive mismatch detection in early infancy. *Neuroreport, 13(10),* 1251-1254.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1456), 815-836.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127-138.

Gagnepain, P., Henson, R.N., & Davis, M.H. (2012). Temporal predictive codes for spoken words in auditory cortex. Current Biology, 22, 615–621.

Gambi, C., Pickering, M. J., & Rabagliati, H. (2016). Beyond Associations: Sensitivity to structure in pre-schoolers' linguistic predictions. *Cognition*, *157*, 340-351.

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science, 15(6),* 297-301.

Herrmann, B., Maess, B., Hasting, A. S., & Friederici, A. D. (2009). Localization of the syntactic mismatch negativity in the temporal cortex: an MEG study. *NeuroImage, 48*(3), 590-600.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118-135.

Ito, A., Martin, A. E., & Nieuwland, M. S. (2016). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience, 32(8)*, 954-965.

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. Psychological Science, 18(3), 193-198.

Lukyanenko, C., & Fisher, C. (2016). Where are the cookies? Two- and three-year-olds use number-marked verbs to anticipate upcoming nouns. *Cognition, 146*, 349-370. doi:http://dx.doi.org/10.1016/j.cognition.2015.10.012

MacWhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. *Journal of Child Language, 31(4)*, 883-914.

Mahr, T., McMillan, B. T., Saffran, J. R., Weismer, S. E., & Edwards, J. (2015). Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition, 142*, 345-350.

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake - but only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance, 38*(4), 843-847.

Marcus, G. F. (1995). Children's overregularization of English plurals: A quantitative analysis. *Journal of Child Language, 22(2)*, 447-459.

Martin, C. D., Thierry, G., Kuipers, J.-R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. Journal of Memory and Language, 69(4), 574-588.

McCauley, S.M. & Christiansen, M.H. (2014a). Acquiring formulaic language: A computational model. *Mental Lexicon, 9,* 419-436.

McCauley, S.M. & Christiansen, M.H. (2014b). Reappraising lexical specificity in children's early syntactic combinations. In P. Bello, M. Guarini, M. McShane & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1000-1005). Austin, TX: Cognitive Science Society.

McKee, C., & McDaniel, D. (2009). Frequency effects in children’s syntactic and morphological development. In W. D. Lewis, S. Karimi, H. Harley, & S. O. Farrar (Eds.), *Time and Again: Theoretical Perspectives on Formal Linguistics: in Honor of*

D. Terence Langendoen (Vol. 135, pp. 177-188). Amsterdam/Philadelphia: John Benjamins.

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Boca Raton, FL: CRC Press.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., . . . Huettig, F. (2017). Limits on prediction in language comprehension: A multi-lab failure to replicate evidence for probabilistic pre-activation of phonology. bioRxiv. doi:10.1101/111807

Oberecker, R., Friedrich, M., & Friederici, A. D. (2005). Neural correlates of syntactic processing in two-year-olds. Journal of Cognitive Neuroscience, 17(10), 1667-1678.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition, 118(3),* 306-338.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(4), 329-392. doi:http://dx.doi.org/10.1017/S0140525X12001495

Rabagliati, H., Gambi, C., & Pickering, M. J. (2016). Learning to predict or predicting to learn?. *Language, Cognition and Neuroscience, 31(1)*, 94-105.

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language, 89*(4), 760-793.

Raymond, W. D., Fisher, J. A., & Healy, A. F. (2002). Linguistic knowledge and language performance in English article variant preference. *Language and Cognitive Processes, 17*(6), 613-662.

Snedeker, J. (2013). Children's sentence processing. In R. P. G. Van Gompel (Ed.), Sentence Processing (pp. 189-220). Hove: Psychology Press.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49(3),* 238-299.

Snedeker, J., & Yuan, S. (2008). Effects of prosodic and lexical constraints on parsing in young children (and adults). *Journal of Memory and Language, 58(2),* 574-608.

Trueswell, J. C. (2008). Using eye-movements as a developmental measure within psycholinguistics. In I. A. Sekerina, E. M. Fernandez, & H. Clahsen (Eds.), *Language processing in children* (pp. 73-96). Amsterdam: John Benjamin.

Van Heugten, M., & Shi, R. (2009). French-learning toddlers use gender information on determiners during word recognition. Developmental Science, 12(3), 419-425.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67(6),* 1176-1190.

Yan, S., Kuperberg, G. R., & Jaeger, T.F. (2017). Prediction (or not) during language processing. A commentary on Nieuwland et al. (2017) and DeLong et al. (2005). bioRxiv. doi: 10.1101/143750

Ylinen, S., Bosseler, A., Junttila, K., & Huotilainen, M. (2016). Predictive coding accelerates word recognition and learning in the early stages of language development. *Developmental Science*, DOI: 10.1111/desc.12472.

Supplementary Material for

Gambi, C., Gorrie, F., Pickering, M.J., & Rabagliati, H. (2018). *The development of linguistic prediction: Predictions of sound and meaning in 2-to-5 year olds.*

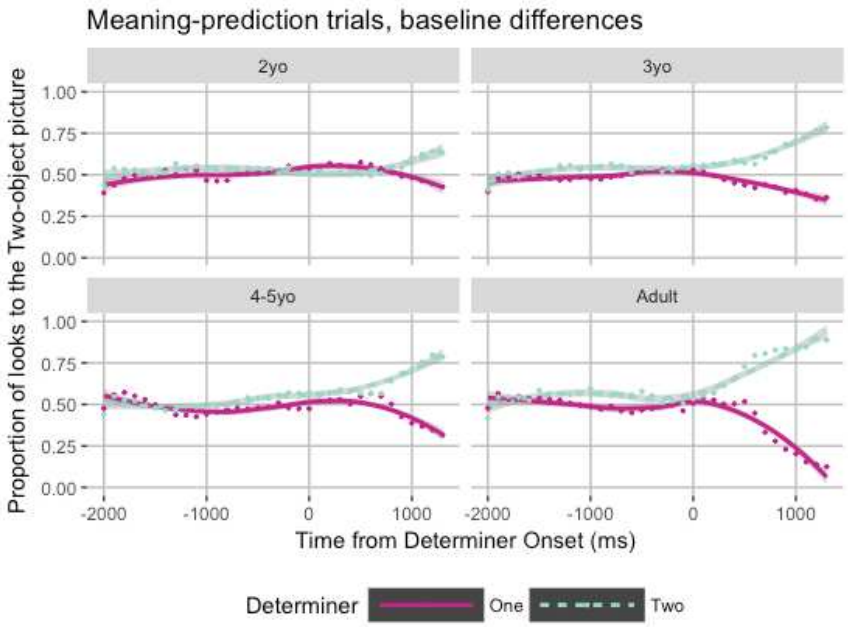
List of Visual World Studies with Children Using Repeated Presentation of Pictures

In the main paper, we report that children learned the repeated trial structure of our experiment. To assess whether this may be a problem for previous visual-world studies that have looked at children's processing of determiners, we conducted a small survey of the word recognition literature cited in the main paper, that focused on the presence and number of picture repetitions, and whether pictures were yoked (i.e., always presented in the same pairs). Repetition of pictures was common. Johnson (2005) used 4 pictures, each of which appeared 6 times; Lew-Williams and Fernald (2007) used 8 pictures, each repeated 8 times; Van Heugten and Shi (2009) used 4 pictures, each repeated 6 times; Melançon and Shi (2015) used two pictures of novel objects that were each repated 6 times, and 5 pictures of familiar objects which were each repeated twice; Mahr et al. (2015) used 16 pictures, each repeated twice, although they corresponded to only 4 different entities (so each entity was repeated 8 times). It was less common that studies explicitly reported whether pictures were yoked. Mahr et al. (2015) and Melançon and Shi (2015; for novel objects only) report that their pictures were yoked. Van Heugten and Shi (2009) list all of their trials: pictures pairs were repeated across grammatical trials where picture had different genders and ungrammatical trials, but were re-yoked to create grammatical trials where pictures had the same gender. It is not clear whether pictures were yoked or not in the other studies, but it is clear that the same picture served both as a target and a distractor on different trials.

Post-hoc Analyses of Experiment 1

Baseline differences.

Figure S1 – Proportion of looks to the two-object picture (top) or to the vowel-initial picture (bottom) over time, starting from 2 seconds before determiner onset. Points are observed values; lines were obtained by non-parametric smoothing (method="loess", with 95% confidence intervals shaded in grey).



First vs. second occurrence of picture pairs.

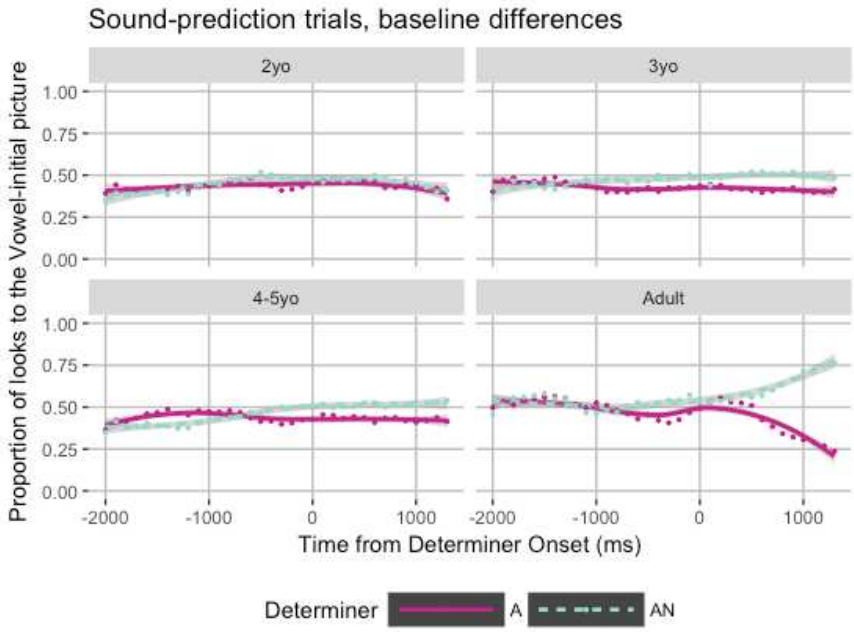


Figure S2 – Meaning-prediction trials. A breakdown of Figure S1 by picture occurrence (First or Second).

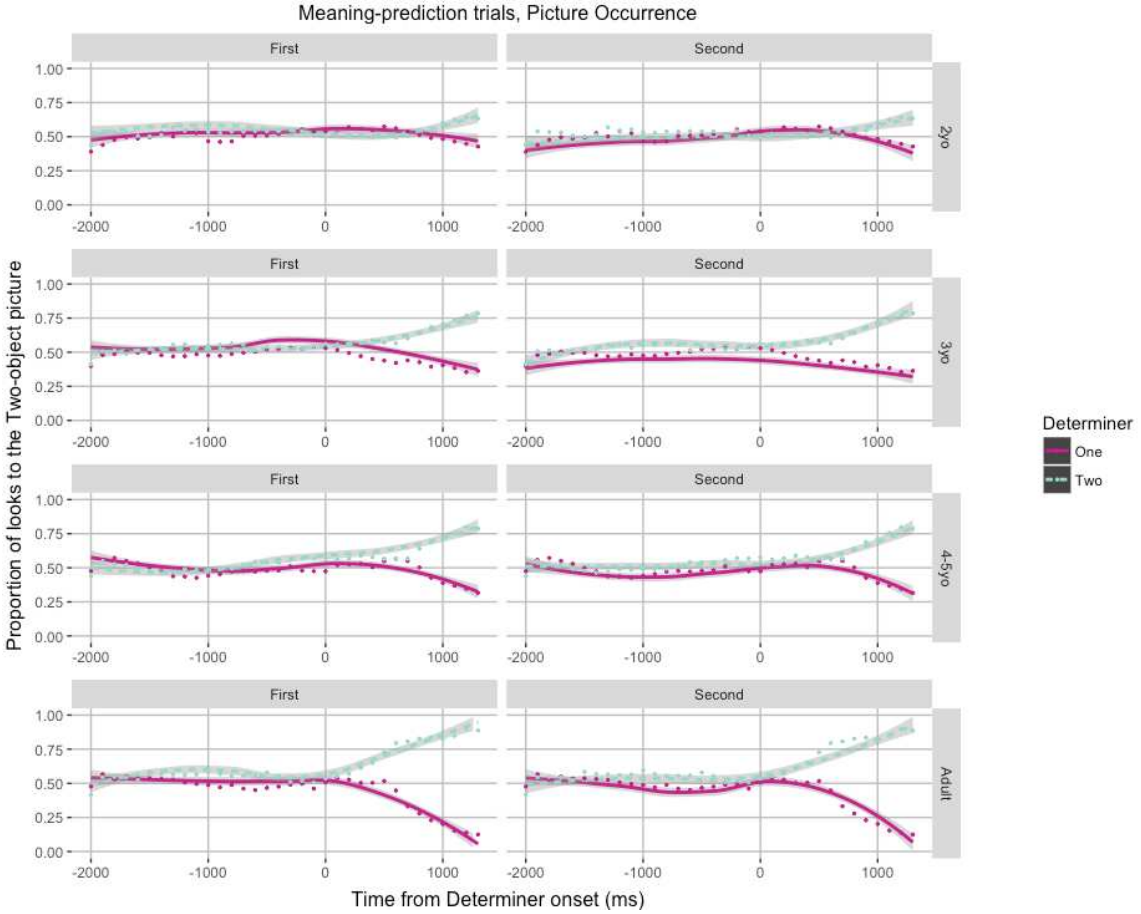
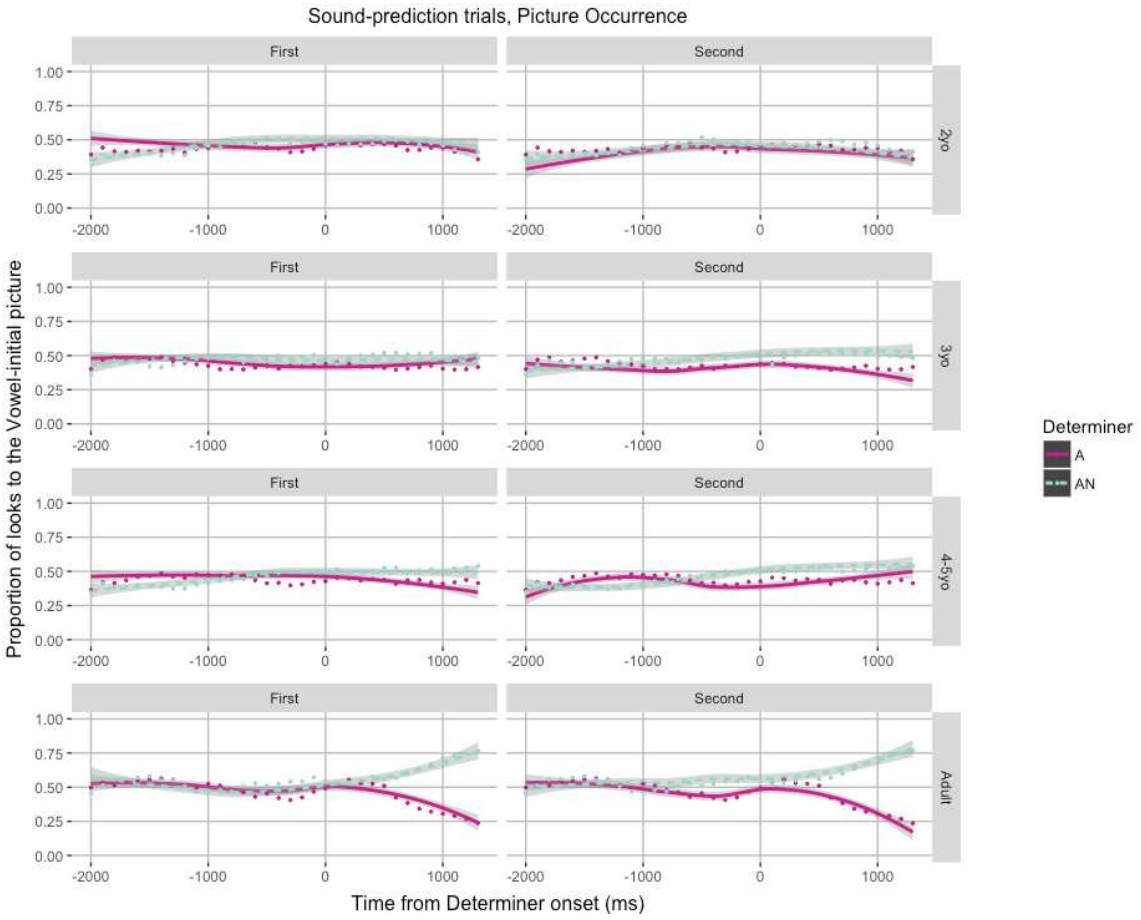


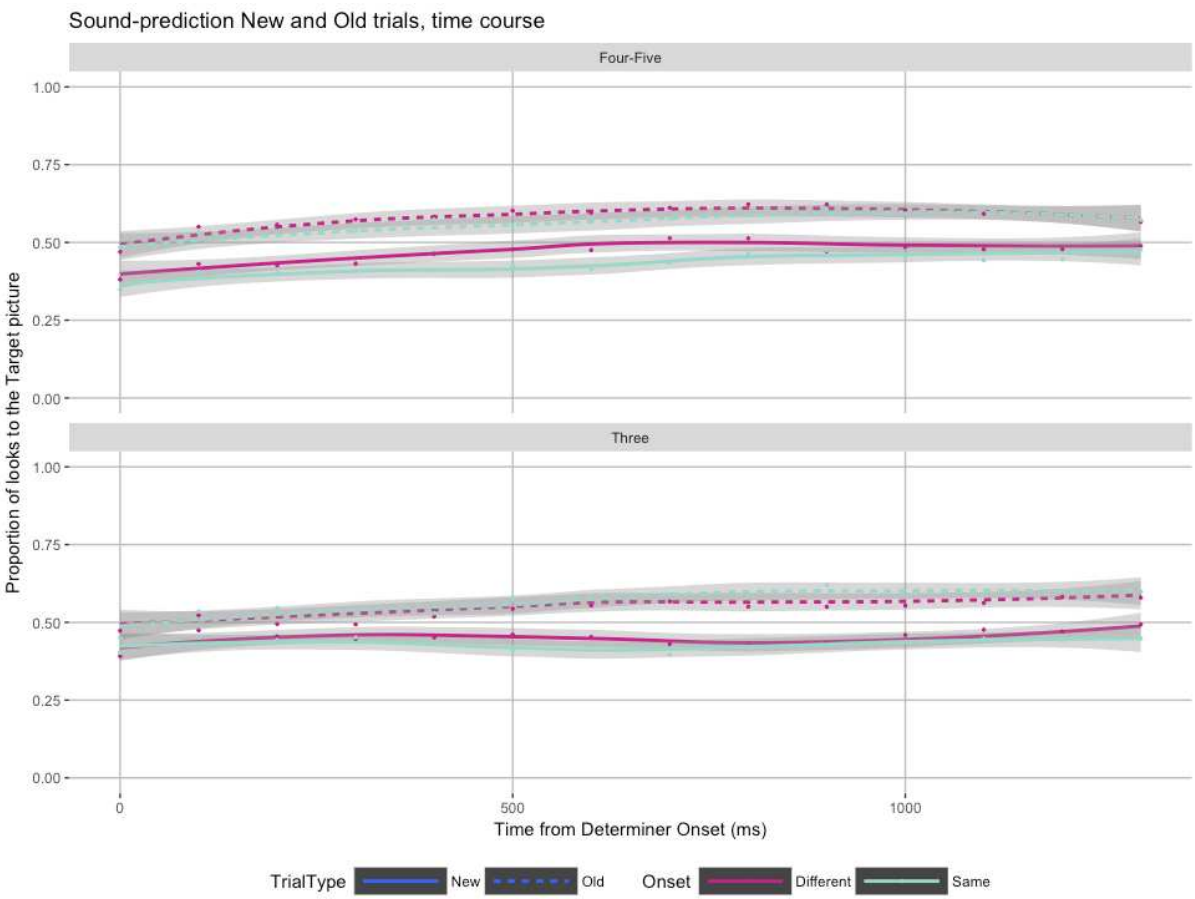
Figure S3 – Sound-prediction trials. A breakdown of Figure S1 by picture occurrence (First or Second)



Experiment 2. Time Course Graphs.

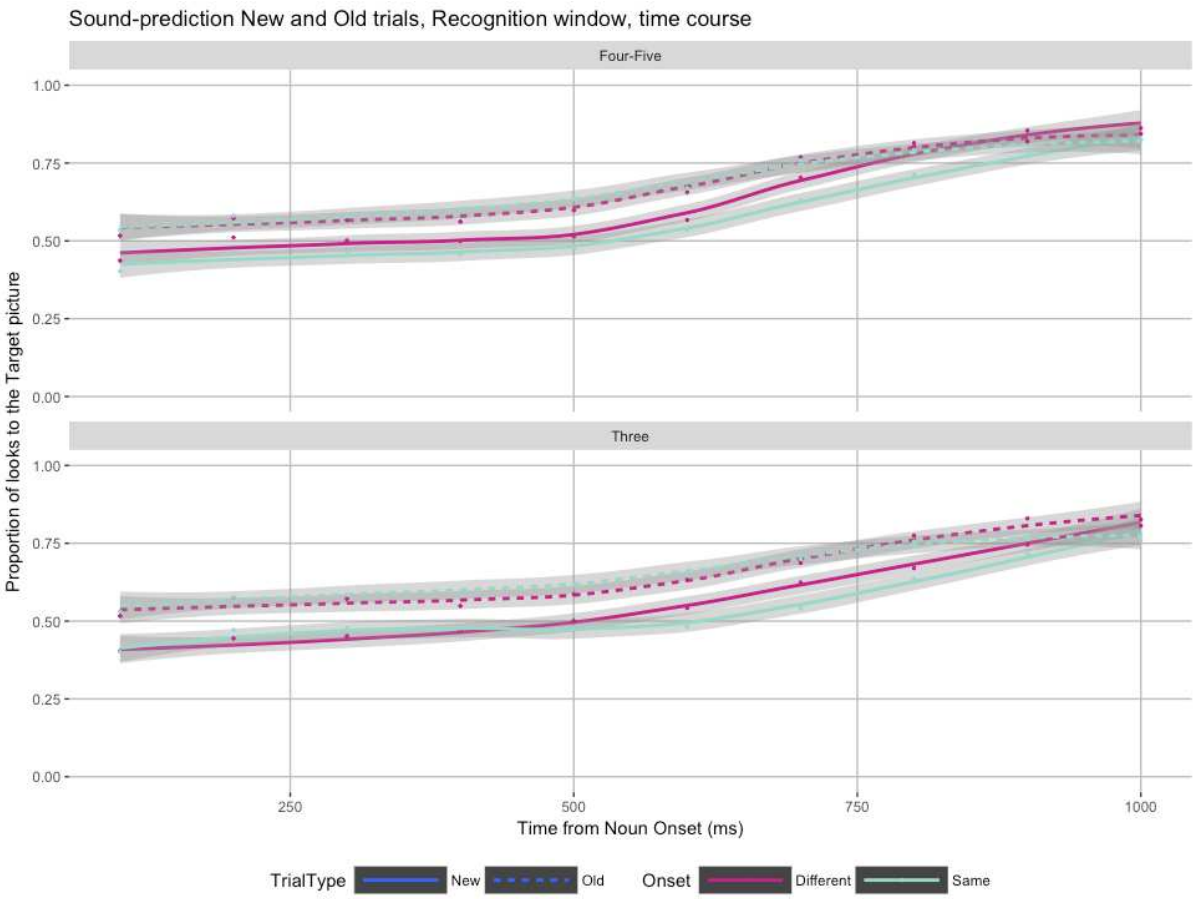
Prediction window.

Figure S4. Sound-prediction New and Old trials. Prediction window. Average proportion of looks to the target picture on Different Onset (pink lines) and Same Onset trials (blue lines). Solid lines are for New Trials, dashed lines for Old trials. Points are observed values; lines were obtained by non-parametric smoothing.



Recognition window.

Figure S5. Sound-prediction New and Old trials. Recognition window. All details are as in Figure S4.



References

Johnson, E. K. (2005). Grammatical gender and early word recognition in Dutch. In: A.Brugos, M. R.Clark-Cotton, & S.Ha (Eds.), *BUCLD 29: Proceedings of the 29th Annual Boston University Conference on Language Development*, Volume 1 (pp. 320–330). Somerville, MA: Cascadilla.

Van Heugten, M., & Shi, R. (2009). French-learning toddlers use gender information on determiners during word recognition. *Developmental Science, 12(3),* 419-425.

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science, 18(3),* 193-198.

Mahr, T., McMillan, B. T., Saffran, J. R., Weismer, S. E., & Edwards, J. (2015). Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition, 142*, 345-350.

Melançon, A., & Shi, R. (2015). Representations of abstract grammatical feature agreement in young children. *Journal of child language, 42(6),* 1379-1393.