Infants' developing sensitivity to native language phonotactics: A Meta-analysis

Megha Sundara, Z.L. Zhou, Canaan Breiss, Hironori Katsuda, & Jeremy Steffman*

UCLA Department of Linguistics

*Currently at Northwestern University

Corresponding Author

Megha Sundara

Department of Linguistics

UCLA

3125 Campbell Hall

Los Angeles

megha.sundara@humnet.ucla.edu

Abstract

We used Bayesian modeling to aggregate experiments investigating infants' sensitivity to native language phonotactics. Our findings were based on data from 84 experiments on about 2000 infants, learning 8 languages tested using 4 different methods. Our results showed that, unlike with artificial languages, infants do exhibit sensitivity to native language phonotactic patterns in a lab setting. However, the exact developmental trajectory depends on the phonotactic pattern being tested. Before 8 months, infants tuned into non-local dependencies between vowels: specifically, vowel harmony. Between 8- and 10-months, infants demonstrated a consistent sensitivity to both local and non-local consonant dependencies. Sensitivity to non-local vowel dependencies that are not based on harmony emerged only after 10-months. These findings provide a benchmark for future experimental and computational research on the acquisition of phonotactics.

1 Introduction

From research over the last several decades, we know that infants become sensitive to their native language speech sound categories during the first year of life. Within that year, infants tune into the set of sounds that are contrastive in any given language that they are acquiring as well as the properties of individual speech sound categories. In this paper, we present a meta-analysis of cross-linguistic research on when and what infants know about the phonotactics — the position, sequencing, and frequency of sounds — in their native language.

Infants' sensitivity to phonotactics is typically indexed by a differential response to more versus less frequent sounds or sequences of their native language. The differential behavioral response can include, depending on the age of the infants and the experimental paradigm, listening or looking preference, successful discrimination, segmentation, and even word learning. Because infants have exposure to their native language in their daily lives, they are expected to preferentially attend to, segment, and learn words with the more frequent patterns if they are sensitive to native language phonotactics.

The now-classic finding by Jusczyk, Luce, & Charles-Luce (1994) set the stage for the research program on the acquisition of phonotactics. Jusczyk et al. presented lists of nonce monosyllables that had high or low probability sound patterns from English to English-learning 6- and 9-month-olds. Infants were tested using the headturn preference procedure. English-learning 9-, but not 6-month-olds listened significantly longer to the high probability lists, leading Jusczyk et al. to conclude that infants are tuning into the phonotactic patterns of their native language between 6 and 9 months. This developmental timeline continues to be the established wisdom in the literature.

Following Jusczyk et al., the bulk of the research on infants' sensitivity to phonotactics has been focused on 9-month-olds. Specifically, we now have converging evidence that English-learning 9-month-olds have knowledge of sequencing restrictions on consonant clusters (CC) in tasks measuring preference (e.g., Archer & Curtin, 2011; Mattys et al., 1999) and word segmentation (e.g., Archer & Curtin, 2016; Mattys & Jusczyk, 2001). Similar significant listening preferences for high probability CC sequences of their native language have also been reported for 9-month-olds learning Dutch (Friederici & Wessels, 1993) and Catalan (Sebastian-Galles & Bosch, 2002).

Older infants also demonstrate knowledge of native language phonotactics. English-learning infants continue to favor high probability CC sequences when learning words in the second year of life (e.g., MacKenzie, Curtin & Graham, 2012; Graf Estes, 2014; Graf Estes, Edwards & Saffran, 2011); Japanese-learning 12- and 18-month-olds successfully discriminate high, but not low probability CC(V) sequences (Kajikawa et al., 2006; Mugitani et al., 2007); French-learning 10-month-olds demonstrate knowledge of non-adjacent dependencies between consonants in tasks measuring preference (e.g., Gonzalez-Gomez & Nazzi, 2012a; Nazzi et al., 2009), segmentation (Gonzalez-Gomez & Nazzi, 2013), and word learning (Gonzalez-Gomez, Poltrock, & Nazzi, 2013). Hebrew- (Segal, Keren-Portnoy & Vihman, 2015) and Turkish-learning 10-month-olds, too, demonstrate sensitivity to language-specific restrictions on non-adjacent dependencies, but between vowels (Altan, Kaya, & Hohenberger, 2016; Hohenberger, Altan, Kaya, Köksal-Tuncer & Avcu, 2016; Hohenberger, Kaya & Altan, 2017). These findings with a variety of phonotactic patterns are consistent with an account where infants are tuning into native language phonotactics by about 9 months, though even the older infants' behavior might not be otherwise adult-like.

The results from research on infants younger than 9-months, however, are both more limited and more equivocal. Although English-learning 6-month-olds fail to demonstrate sensitivity to sequencing restrictions on CC clusters in a word segmentation task (Mattys et al., 1999), they do so in a preference task (Archer & Curtin, 2011). Yet, even in preference tasks, Dutch-learning 4.5- and 6-month-olds fail to favor high probability CC clusters (Friederici & Wessels, 1993). Research on French-learning infants also shows a discrepancy in the performance of younger infants depending on the specific phonotactic pattern being tested: French-learning 7-month-olds prefer listening to sequences of more frequent segments, but not sequences with more frequent non-adjacent consonant sequences (Gonzalez-Gomez & Nazzi, 2012a). Like the French-learning infants, Japanese-learning 7-month-olds also fail to show a preference for sequences with more frequent non-adjacent consonant sequences (Gonzalez-Gomez, Hayashi, Tsuji, Mazuka & Nazzi, 2014). However, Turkish-learning 6-month-olds do prefer more frequent non-adjacent sequences involving vowels (Altan, Kaya, & Hohenberger, 2016; Hohenberger, Altan, Kaya, Köksal-Tuncer & Avcu, 2016; Van Kampen, Parmaksiz, Van De Vijver, & Hohle, 2008; see also Hohenberger, Kaya, & Altan, 2017 for an asymmetry in discrimination consistent with a preference for harmonic sequences).

To summarize, there is converging evidence that infants learning many different languages are sensitive to phonotactic restrictions at 9-months, as are older infants. The most variability, however, is in the performance of younger infants: in infants younger than 9-months, whether or not studies demonstrate sensitivity to phonotactics seems to depend on the task as well as the specific pattern being tested in addition to the infants' native language.

In the first part of the meta-analysis reported in this paper, we addressed whether infants are sensitive to native language phonotactics. This is a non-trivial question considering the climate

of concern about replicability, underpowered studies in infant research (Oakes, 2017), as well as a recent meta-analytic report calling into question that infants can learn phonotactics just from short-term exposure to artificial languages in the lab (Cristia, 2018). In the second part, we evaluated the extent to which infants' sensitivity to phonotactics is moderated by age, testing methods, and phonotactic dependency type.

To establish a developmental timeline we need to be sure at what ages infants fail and succeed on a given task. However, as Bergmann et al. (2018) review, the interpretation of nonsignificant findings is fraught: researchers cannot be sure whether any particular null finding is a true lack of ability, noise, or an error because of an insensitive experimental design or small sample size. Throughout the paper, we used Bayesian models to integrate evidence from data (here, effect sizes from comparisons of more and less frequent items in a paper) with prior expectations, to quantify the strength of evidence for likely values of parameters of interest. This approach is particularly well-suited to meta-analysis because it allows the opportunity to quantify the contribution of individual studies to our estimate about the object of inference — our certainty about infants' degree of sensitivity to phonotactic patterns — rather than having to rely on a binary accept/reject criterion evaluating a null hypothesis of an effect size of zero (Kruschke, 2014). Bayesian inference also allows for principled synthesis of continuous levels of evidence provided by "significant" as well as "non-significant" findings in the summarized literature (Nicenboim et al., 2018). Using Bayesian modeling in this meta-analysis, we focused on establishing the age at which infants tune into specific phonotactic patterns in their native language.

2 Methods

2.1. Paper identification and selection

The term ‘paper’ is used here as a cover term for conference proceedings, published journal articles, book chapters, and unpublished reports. The initial cohort was compiled by identifying papers known to the authors (26 papers) and by systematic searches of databases and reference lists (195 papers).

The 221 paper titles and abstracts were then screened. After removing duplicates (38 papers), we excluded papers which did not meet the following broad selection criteria. First, in order to be included in the meta-analysis, a paper needed to examine learning of segmental phonotactics in infants under the age of 2 years in their native language. Second, only papers using behavioral measures such as the headturn preference procedure, central fixation, and preferential looking paradigm were included. Third, we did not include papers which tested learning of artificial or modified languages. Because discrimination experiments habituate infants in the lab to one or other stimuli with a view to affecting behavioral outcomes, we treated such experiments as involving artificial languages and excluded them (5 papers).

After the initial screening, we excluded papers which tested adult phonotactic learning (25), those that were not about the learning of phonotactics (31), involved prosodic/suprasegmental features (14), used artificial languages (8), were computational simulations (5), were studies of corpora (2), tested children older than two years (3), were review articles which did not present new experimental data (12), used methodologies which did not fit our criteria (7), and for which we were unable to obtain the text of the paper (1). This left us with 70 papers.

We examined the full text of the remaining 70 papers to determine their eligibility for the meta-analysis. This eligibility assessment resulted in the additional exclusion of papers for not being about phonotactics (29), testing adults (2), testing phonotactic patterns unrelated to infants’

native language (1), testing only sequences which were illegal in the infants' native language (1), testing children over the age of two (5), lacking critical information needed for the meta-analysis (2), and for providing only a subset of data included in another paper (1), or not providing any data (2). After screening, a total of 27 papers were eligible for the meta-analysis (see PRISMA flowchart in supplementary material on the OSF page for this project https://osf.io/ecpjz/?view_only=1c45fcf9a72848b5bb6dd98552c8bd85).

2.2 Data entry

Papers in the final cohort were entered into the analysis. Some papers included multiple experiments or conditions, in which case each comparison constituted a record of an experiment for data entry, if it provided an independent effect size estimate. For example, Van Kampen et al. (2008) tested infants on two sets of stimuli where one set had initial stress and the other had final; because they compared high and low frequency items within each set, this paper contributed two entries to our analysis.

Each record was coded for a number of dimensions following previous meta-analyses (Bergman & Cristia, 2016; Cristia, 2018; for a full list and explanation of coded variables see Bergman et al., 2018). The relevant dimensions for the present analysis were, (1) background information on the paper, including year of publication, and whether the publication was peer reviewed; (2) the number of infants recruited and the number of infants excluded; (3) infants' native language; (4) infants' mean age and age range in days; (5) whether a trial type was more or less frequent in the infant's native language; and (6) the mean and standard deviation of looking times for each trial type. For 3 papers (4 experiments; Graf Estes, 2014; Graf Estes & Bowen, 2013; MacKenzie, Curtin, & Graham, 2012) we derived a difference measure between looking times to same and switch trials for high probability and low probability conditions separately and

used those as looking time measures for high and low frequency trial types. We additionally coded for moderator variables describing the type of phonotactic pattern tested: (1) whether the tested patterns occurred at word edges as compared to word medially, and (2) whether low frequency items in a study obeyed sonority sequencing principles.

The final dataset had results from 1,924 infants learning 8 languages, using 4 methods in 84 experiments from 27 studies. The distribution of age and native language of the infant in the final sample are presented in Figure 1.

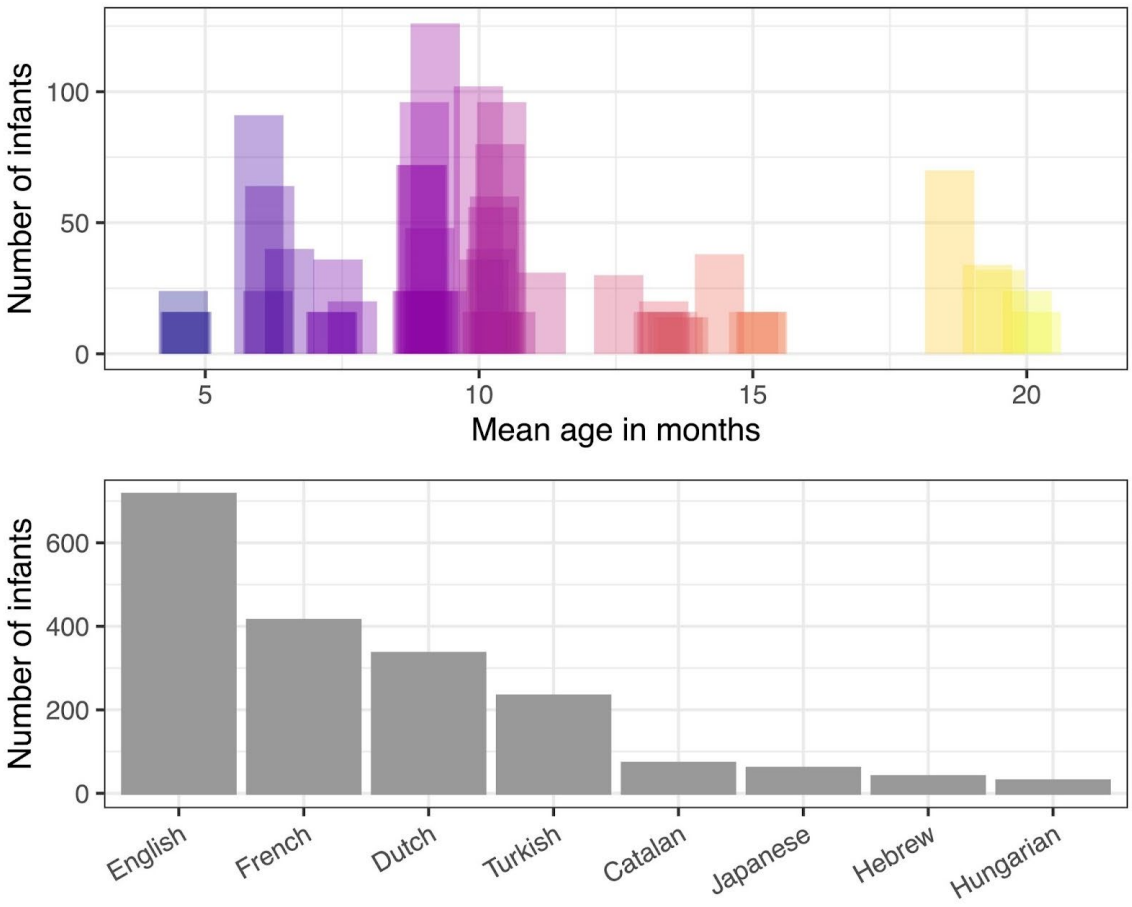

Figure 1. Distribution of the number of infants in the meta-analysis based on age (for a given experiment, top), and based on native language (bottom).

2.3 Derived variables

2.3.1 Effect size

In order to compare the strength of evidence across studies, we converted infants' response to the high and low frequency pattern within each study to a standardized effect size. We used the *esc_mean_sd()* function from the *esc* package to calculate effect sizes (Lüdecke, 2019). Effect size is commonly defined as the difference between two sample means, divided by the pooled standard deviation of the two means (Cohen's *d*). However, in this meta-analysis, instead of using Cohen's *d*, we used Hedges' *g* because it additionally scales the calculation based on the sample size of each study; as a result, studies with more participants have a proportionally greater impact on the meta-analytic estimate (Shadis & Haddock, 2009). Outside of this, values of Cohen's *d* and Hedges' *g* are interpreted similarly. Hedge's *g* was the dependent variable throughout this paper.

If papers had multiple comparisons involving more and less frequent sequences, we calculated the effect size for each pairwise comparison independently. In cases where the raw data was made available to us, we carried out paired *t*-tests and used this measure instead of the reported omnibus comparison statistic reported in the paper. In these cases, we were also able to incorporate the within-individual correlation between trials into our calculation of effect size, increasing the precision of our estimates.

2.3.2 Standard deviation

A few studies did not report standard deviations (or standard errors) in the text or in figures, but did report means, *n*'s, and *t*-statistics. In such cases, we derived the standard deviation by leveraging the fact that the *t*-statistic is itself derived from the standard deviation of the difference between means, and that the standard deviation of a condition is equal to the standard deviation of

the difference between conditions, if the conditions have the same variance. We confirmed this using a two-tailed paired *t*-test conducted on the standard deviations that were reported in the publications ($t(108)=1$, $p=0.319$); so we included the derived standard deviations in our models.

2.4 Analyses

All statistical analyses reported here were carried out in the R programming environment (R Core Team, 2021), using the *brms* package (Bürkner, 2017, 2018; v. 2.14.4) to fit Bayesian (hierarchical) models to the data. All models included a random intercept of experimental comparison nested within paper to account for the residual variance arising from possibly non-uniform influence of the specific language, testing method, research team, and population being studied on experimental outcomes. Further, all models reported here took into account the uncertainty in the effect size in the original paper: instead of modelling effect size directly, we modeled the effect size drawn from a Normal distribution parameterized by the mean and standard error of Hedge's *g*, as derived described in Section 2.1.1. We report the median value of the posterior distribution for each parameter of interest, along with values denoting the upper and lower limits of an interval that contains the central 95% of values for the parameter (the 95% Credible Interval). From this statistic we can make inferences about the most likely values of the parameter (those that are closer to the median value), and those which are less likely (those closer to the tails of the distribution). For intercepts and coefficients we also report the posterior probability of a positive effect, which ranges between 0 and 1; this statistic is obtained by simply examining the proportion of credible values which lie above zero, and represents the probability of *some* nonzero positive effect, regardless of magnitude.

All models were fit using a No U-Turn Sampler to draw 10,000 samples in each of four Markov chains from the posterior distribution over parameter values, conditioned on the observed

data and our priors. In order to ensure adequate independence from the starting values of the Markov chains, we discarded the first 1,000 samples from each chain, retaining the latter 90% of samples for inference.

Since continuous variables were centered and scaled, we used a Normal(μ=1,σ=0.5) prior on the intercept for the primary analysis reported in Part I; moderator analyses were fit so as to exclude an intercept term. For the Bayesian implementation of Egger's test reported in Section 3.2.1, we used a Normal(0, 2) prior, since we didn't have strong assumptions about the skewness of the funnel plot in question. We used a Normal(0,0.5) prior on coefficient parameters throughout, and a Normal(0,1) prior on standard deviations. These priors were chosen because we had no cause to believe we would find large effect sizes (i.e., $|g| > 1$), but were rather certain that an effect would be found, given that all typically-developing children must learn phonotactics eventually. All $\hat{R}$ values were within 0.01 points of 1, indicating that the Markov chains explored the posterior in an unbiased manner (i.e., the model converged). Posterior samples from the models described here can be obtained from the OSF page for this project. A sensitivity analysis showed little variation in posterior values in the primary analysis for a range of prior expectations including Normal(0,10) and Normal(1,1); further data is available on the OSF page.

3 Results and Discussion

Part I: Are infants sensitive to native language phonotactics?

In this section, we focused on whether infants are sensitive to language specific phonotactics of their native language. Sensitivity to phonotactics during development has been

evaluated in one of two ways. In one approach, infants are tested on patterns from their native language to which they have had long-term exposure commensurate with their age. Alternatively, there are also experiments on what infants may learn from short-term exposure to an artificial language created in the lab. While the greater control and flexibility in designing artificial languages promises the potential to inform about plausible mechanisms available to infants during acquisition, a meta-analysis by Cristia (2018) summarizing this literature found that the overall effect size for such studies was not significantly different from zero. That is, Cristia found no evidence that infants can learn phonotactic patterns only from short term exposure to an artificial language in the lab.

The analyses presented in Part I were targeted towards comparing infants' sensitivity to phonotactic patterns learned from long-term native language exposure vs. short-term lab experiments. First, we report results from a Bayesian model evaluating the effect size of experiments where infants were tested on native language patterns to which they had long-term exposure. Then, we re-analyzed Cristia's data (2018) from artificial language experiments using Bayesian models for an integrative analysis in which we fitted a combined model to both datasets, with Language (*artificial* vs. *natural*) as a fixed effect. If infants can learn phonotactics from native language input within the first two years, we expect a positive effect size with natural language stimuli. This finding would be additionally strengthened in case of a significant effect of Language in the joint analysis of data reported here and in Cristia (2018).

3.1 Infants do learn native language phonotactics

To address whether infants are sensitive to native language phonotactics we calculated the aggregate meta-analytic effect size across all experiments. To this end, we fit a random-effects-only meta-analytic measurement-error model. In this model, the effect sizes for individual

comparisons in each experiment can be thought of as standing in for the "population" of individuals in a more traditional random-effects experimental design. The model also incorporates the uncertainty in each study's estimate of the true effect size by modeling the point-estimate as sampled from a normal distribution parameterized by the standard error of the effect size for the study in question. This allowed us to retain the uncertainty in the estimates from each of the experiments included rather than assuming perfect accuracy of measurement, thereby reducing potentially overconfident estimates of our model. The intercept in this model is the statistical and scientific quantity of interest: it represents the population-level estimate of the effect size associated with infant sensitivity to native language phonotactics as assessed by all the literature we gathered.

The meta-analytic effect size across all experiments was 0.38, with 95% Credible Interval (CrI)=[0.20, 0.57], $p(\beta > 0) = 1.00$. That is, based on the results from all experiments included here, we can be 100% certain that the effect size is positive, and 95% confident that the effect size is between 0.20 and 0.57. Thus, aggregating across age, pattern, language, and experimental method, infants showed a medium-small, consistent preference for more frequent sequences from their native language. Figure 2 displays the meta-analytic effect size (bottom), along with the shrunken estimates for the median and 95% CrI of the effect size of interest from each of the studies included in the meta-analysis.


Graf.Estes.et.al..(2011)_1_Older_than_10mos_36
MacKenzie.et.al..(2012a)_1_Older_than_10mos_112
Atlan.et.al..(2016)_1_Below_8mos_94
Atlan.et.al..(2016)_2_Below_8mos_95
Gonzalez-Gomez.& Nazzi.(2016)_1_Older_than_10mos_65
Gonzalez-Gomez.& Nazzi.(2016)_2_Older_than_10mos_67
Gonzalez-Gomez.& Nazzi.(2012b)_1_Between_8_and_10mos_40
Jusczyk.et.al..(1993)_1_Between_8_and_10mos_4
Gonzalez-Gomez.(2012d)_3_Between_8_and_10mos_108
Gonzalez-Gomez.& Nazzi.(2012a)_1_Between_8_and_10mos_75
Jusczyk.et.al..(1993)_4_Between_8_and_10mos_3
Jusczyk.et.al..(1993)_4_Between_8_and_10mos_5
Friederici.& Wessels.(1993)_1_Between_8_and_10mos_12
Friederici.& Wessels.(1993)_6_Between_8_and_10mos_11
Jusczyk.et.al..(1994)_1_Between_8_and_10mos_20
Gonzalez-Gomez.& Nazzi.(2012c)_1_Between_8_and_10mos_46
Nazzi.et.al..(2009)_1_Between_8_and_10mos_35
Gonzalez-Gomez.& Nazzi.(2012c)_2_Between_8_and_10mos_49
Segal.et.al..(2015)_2_Between_8_and_10mos_91
Gonzalez-Gomez.et.al.(2014)_1_Older_than_10mos_73
Segal.et.al..(2015)_1_Between_8_and_10mos_89
Friederici.& Wessels.(1993)_5_Between_8_and_10mos_10
Gonzalez-Gomez.(2012d)_4_Between_8_and_10mos_111
Jusczyk.et.al..(2002)_1_Between_8_and_10mos_81
Jusczyk.et.al..(2002)_6_Below_8mos_80
Mattys.et.al..(1999)_1_Between_8_and_10mos_24
Gonzalez-Gomez
Mattys.et.al..(1999)_3_Between_8_and_10mos_22
Mattys.& Jusczyk.(2001)_2_Between_8_and_10mos_26
Jusczyk.et.al..(2002)_4_Below_8mos_79
Graf.Estes.& Bowen.(2013)_1_Older_than_10mos_43
Mattys.& Jusczyk.(2001)_1_Between_8_and_10mos_27
Jusczyk.et.al..(2002)_3_Between_8_and_10mos_82
Sebastián-Gallés.& Bosch.(2002)_1_Between_8_and_10mos_31
Jusczyk.et.al..(2002)_8_Older_than_10mos_85
Mattys.& Jusczyk.(2001)_3_Between_8_and_10mos_25
Jusczyk.et.al..(1994)_3_Between_8_and_10mos_21
Mattys.et.al..(1999)_2_Between_8_and_10mos_23
Graf.Estes.2014_2_Older_than_10mos_45
Jusczyk.et.al..(2002)_10_Older_than_10mos_87
Zamuner.(2013)_1_Older_than_10mos_78
Sebastián-Gallés.& Bosch.(2002)_2_Between_8_and_10mos_30
Van.Kampen.et.al..(2008)_1_Below_8mos_32
Jusczyk.et.al..(1999)_4_Between_8_and_10mos_107
Van.Kampen.et.al..(2008)_1_Below_8mos_33
Gonzalez-Gomez.et.al..(2019)_2_Older_than_10mos_100
Friederici.& Wessels.(1993)_3_Between_8_and_10mos_16
Gonzalez-Gomez.& Nazzi.(2016)_2_Between_8_and_10mos_63
Gonzalez-Gomez.& Nazzi.(2012b)_1_Below_8mos_38
Gonzalez-Gomez.et.al.(2014)_1_Between_8_and_10mos_71
Friederici.& Wessels.(1993)_1_Below_8mos_6
Nazzi.& Bertoncini.(2009)_1_Older_than_10mos_92
Sebastián-Gallés.& Bosch.(2002)_2_Between_8_and_10mos_28
Gonzalez-Gomez.et.al..(2019)_1_Between_8_and_10mos_99
Jusczyk.et.al..(1999)_2_Between_8_and_10mos_106
Jusczyk.et.al..(1993)_6_Below_8mos_2
Gonzalez-Gomez.& Nazzi.(2012b)_2_Between_8_and_10mos_41
Jusczyk.et.al..(1999)_1_Between_8_and_10mos_102
Friederici.& Wessels.(1993)_4_Between_8_and_10mos_13
Nazzi.& Bertoncini.(2009)_1_Older_than_10mos_93
Gonzalez-Gomez.et.al.(2014)_1_Below_8mos_70
Jusczyk.et.al..(1993)_3_Below_8mos_1
Jusczyk.et.al..(2002)_7_Older_than_10mos_83
Jusczyk.et.al..(2002)_9_Older_than_10mos_86
Gonzalez-Gomez.& Nazzi.(2016)_1_Between_8_and_10mos_62
Friederici.& Wessels.(1993)_2_Between_8_and_10mos_15
Nazzi.et.al..(2009)_1_Below_8mos_34
Gonzalez-Gomez.& Nazzi.(2016)_3_Older_than_10mos_64
Zamuner.(2006)_1_Between_8_and_10mos_76
Zamuner.(2006)_1_Older_than_10mos_77
Gonzalez-Gomez.& Nazzi.(2016)_3_Older_than_10mos_66
Graf.Estes.& Bowen.(2013)_1_Older_than_10mos_44
Gonzalez-Gomez.& Nazzi.(2012a)_1_Below_8mos_74
Jusczyk.et.al..(1994)_2_Below_8mos_18
Gonzalez-Gomez.& Nazzi.(2012b)_2_Between_8_and_10mos_42
Friederici.& Wessels.(1993)_1_Below_8mos_9
Friederici.& Wessels.(1993)_2_Between_8_and_10mos_14
Gonzalez-Gomez.& Nazzi.(2012b)_3_Below_8mos_37
Gonzalez-Gomez.(2012d)_3_Between_8_and_10mos_109
Gonzalez-Gomez.& Nazzi.(2012b)_3_Below_8mos_39
Gonzalez-Gomez.et.al..(2019)_1_Older_than_10mos_101
Gonzalez-Gomez.(2012d)_4_Between_8_and_10mos_110
Atlan.et.al..(2016)_2_Between_8_and_10mos_98
Atlan.et.al..(2016)_1_Between_8_and_10mos_97
Pooled Effect
2.34 [1.56, 3.02]
2.25 [1.25, 3.31]
1.98 [1.47, 2.51]
1.97 [1.38, 2.59]
1.09 [0.38, 1.8]
1.05 [0.35, 1.75]
1 [0.31, 1.72]
0.92 [0.31, 1.51]
0.84 [0.17, 1.54]
0.84 [0.22, 1.47]
0.75 [0.16, 1.34]
0.72 [0.11, 1.3]
0.69 [0.12, 1.26]
0.68 [0.12, 1.26]
0.67 [0.09, 1.27]
0.68 [-0.02, 1.36]
0.67 [-0.01, 1.36]
0.66 [-0.03, 1.34]
0.64 [0.01, 1.27]
0.63 [0.01, 1.27]
0.61 [-0.02, 1.25]
0.6 [0.03, 1.2]
0.6 [-0.07, 1.29]
0.59 [-0.08, 1.26]
0.57 [-0.1, 1.23]
0.56 [-0.04, 1.15]
0.53 [-0.14, 1.2]
0.51 [-0.07, 1.1]
0.5 [-0.09, 1.08]
0.5 [-0.16, 1.16]
0.48 [-0.16, 1.14]
0.48 [-0.11, 1.07]
0.48 [-0.17, 1.15]
0.47 [-0.11, 1.06]
0.46 [-0.19, 1.12]
0.46 [-0.13, 1.04]
0.46 [-0.13, 1.04]
0.45 [-0.15, 1.03]
0.44 [-0.2, 1.08]
0.43 [-0.21, 1.1]
0.43 [-0.1, 0.97]
0.43 [-0.15, 1.02]
0.42 [-0.21, 1.04]
0.41 [-0.16, 1.01]
0.41 [-0.21, 1.02]
0.33 [-0.36, 1.04]
0.32 [-0.24, 0.89]
0.29 [-0.36, 0.95]
0.28 [-0.38, 0.97]
0.28 [-0.33, 0.89]
0.26 [-0.29, 0.85]
0.25 [-0.48, 0.99]
0.25 [-0.33, 0.83]
0.16 [-0.5, 0.86]
0.16 [-0.41, 0.75]
0.14 [-0.44, 0.71]
0.13 [-0.52, 0.83]
0.12 [-0.45, 0.71]
0.1 [-0.47, 0.68]
0.09 [-0.63, 0.83]
0.09 [-0.52, 0.71]
0.09 [-0.49, 0.65]
0.08 [-0.58, 0.73]
0.07 [-0.57, 0.73]
0.06 [-0.6, 0.72]
0.04 [-0.51, 0.63]
0.04 [-0.62, 0.7]
0.03 [-0.62, 0.68]
-0.02 [-0.54, 0.56]
-0.02 [-0.54, 0.56]
-0.02 [-0.67, 0.63]
-0.03 [-0.7, 0.64]
-0.09 [-0.7, 0.52]
-0.11 [-0.69, 0.47]
-0.14 [-0.81, 0.57]
-0.27 [-0.83, 0.32]
-0.27 [-0.84, 0.32]
-0.31 [-0.98, 0.4]
-0.47 [-1.15, 0.23]
-0.5 [-1.19, 0.21]
-0.55 [-1.26, 0.19]
-0.58 [-1.27, 0.14]
-1.02 [-1.48, -0.53]
-1.84 [-2.38, -1.29]
0.38 [0.2, 0.57]
-2
0
2
4
Effect size (Hedge's g)

Figure 2. Median and 95% Credible Interval for the shrunk effect size for each experiment. The vertical dashed red line marks an effect size of zero, while the grey vertical lines mark the meta-analytic median effect size plus upper and lower 95% CrIs.

3.2 Evidence of (slight) bias in the literature

If there is preferential submission or publication of positive results, the meta-analytic effect size reported here could be potentially overinflated. We evaluated bias in two ways.

3.2.1 Some evidence for selective submission or publication practices

First, to evaluate the possibility of preferential submission of positive results, we used a funnel plot (Figure 3) to plot Hedges' *g* as a function of 1/Standard Error. Funnel plots are useful for detecting publication bias resulting from selective recruitment practices or small-study effects on reporting (Sterne & Egger, 2001; Sterne & Harbord, 2004). Researchers may be disproportionately motivated to gather larger samples when an effect in an intermediate analysis is near-significant and in line with their prior expectations, compared to when it is not significant, or even nearing significance in the opposite direction. This practice can be deduced by gaps in the funnel plot, particularly towards the base of the vertical axis when the power is low. Further, lateral asymmetries in the funnel plot can indicate that researchers were unsuccessful at publishing studies that failed to confirm their prior expectations. Because estimates from statistically-significant but low-power studies are guaranteed to be overestimates of the true effect size (Vasishth et al., 2018), these gaps are more likely near the base of the funnel near zero on the horizontal axis, since researchers are biased against publishing non-significant findings. In the funnel plot in Figure 3, the apparent vertical displacement of the points upwards is driven entirely by one study with a very small sample size but a very large effect size.

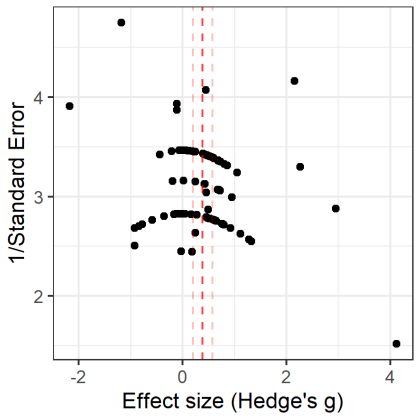

Figure 3. Funnel plot which graphs 1/Standard Error (vertical axis) as a function of Hedge's *g* (on the horizontal axis). Dotted red lines indicate median meta-analytic effect size from model reported in Section 3.1, with pale red lines delimiting the 95% Credible Interval.

While the bulk of the studies were clustered rather symmetrically around the meta-analytic effect size, the outliers were asymmetrically distributed; notably, low-powered studies (low on the vertical axis) tended to skew more to the right on the horizontal axis than to the left, suggesting a possible scenario where studies with comparably small sample sizes were conducted but not published because of non-significant (or indeed, perhaps *opposite*) effects. We confirmed the rightward skew of the overall distribution with a Bayesian implementation of Egger's test (Egger et al. 1997), which regresses the ratio of each study's effect size and its standard error against its inverse standard error. If the estimate for the intercept in this model excludes zero, we can conclude that there is asymmetry in the funnel plot in the direction of the coefficient of the intercept. The results of the test did indicate that there was a slight rightward-skewing asymmetry ($\beta = 3.14$, 95% CrI [0.03, 6.17], $p(|\beta>0) = 0.98$). This suggests that the published literature may be biased in favor of effects which confirmed a priori hypotheses.

3.2.2 No evidence of *p*-hacking

We also carried out a *p*-curve analysis on the collected studies. This is used to evaluate whether researchers have exploited their degrees of freedom (including or excluding specific participants as outliers, trying different statistical tests, different transformations of the dependent variable, etc.) to "*p*-hack" their results.

A *p*-curve analysis examines the distribution of *p*-values below 0.05 to determine whether they are (a) more likely to have arisen from a series of studies testing a robust underlying effect (a right-skewed *p*-curve), (b) indistinguishable from those which would arise under a null underlying effect (flat *p*-curve), or (c) the likely result of extensive *p*-hacking, and therefore of questionable evidential value (left-skewed *p*-curve). Using the *pcurve()* function from the *dmetar* package (Harrer et al. 2019), we found significant evidence of right-skewness in the distribution of *p*-values ($p < 0.001$). The *p*-curve analysis itself had a power of 0.84 (confidence interval 0.71-0.92), indicating that there were enough studies included in the meta-analysis to provide a well-powered robust estimate of the skewness of the *p*-value distribution. This confirms the absence of *p*-hacking in the literature reported. Note that because the test carried out by the *pcurve()* function was not a Bayesian one, we cannot interpret these intervals as *credible intervals*, as has been the general practice in this paper. Because there was no evidence for *p*-hacking and only a slight bias towards selective submission or publication practices, we can be confident regarding the estimates of effect size reported here.

3.3 Comparison to Cristia's (2018) meta-analysis on artificial language learning experiments

Next, we directly compared the meta-analytic review of the 84 experiments evaluated in Section 3.1 with the 34 experiments summarized in Cristia (2018) on teaching infants phonotactics in the laboratory through artificial languages. The results from a Bayesian re-analysis of the

artificial language learning experiments are provided in Appendix 1. The effect sizes for natural vs. artificial languages are summarized in Figure 4.

The difference between the experiments using artificial vs. natural language stimuli was credible. Consistent with the previous analyses, we found that the effect size for artificial languages was very likely near-zero (β = 0.02, 95% CrI=[-0.22, 0.27], $p(\beta>0)$ = 0.55), and the effect size for natural languages was positive, and much larger (β = 0.35, 95% CrI=[0.19, 0.51], $p(\beta>0)$ = 1.00). Note that the estimate obtained here differs slightly from the one provided in Section 3.1 because of the partial pooling among results in the hierarchical model. In the general discussion we present hypotheses to account for infants' success when tested on native language phonotactic patterns and their failure to learn from artificial languages.

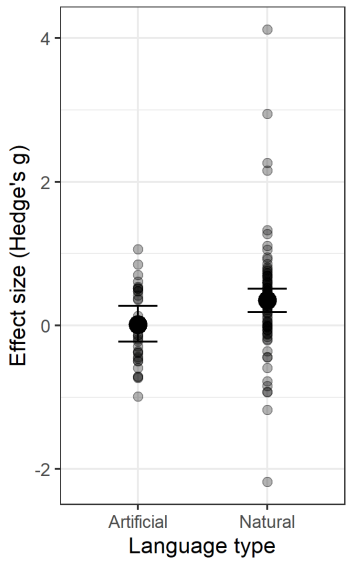

Figure 4. Effect size plotted as a function of language type (*natural* vs. *artificial*). Point estimates are the median of the posterior distribution for each group, with error bars encompassing the 95% Credible Intervals. Each translucent black dot represents an individual effect size from a study in the model.

Part II: Variables influencing infants' sensitivity to native language phonotactics

Given the positive effect size in the developmental literature evaluating sensitivity to native language phonotactics, we investigated which linguistic and methodological variables might moderate the overall effect.

3.4 Method matters

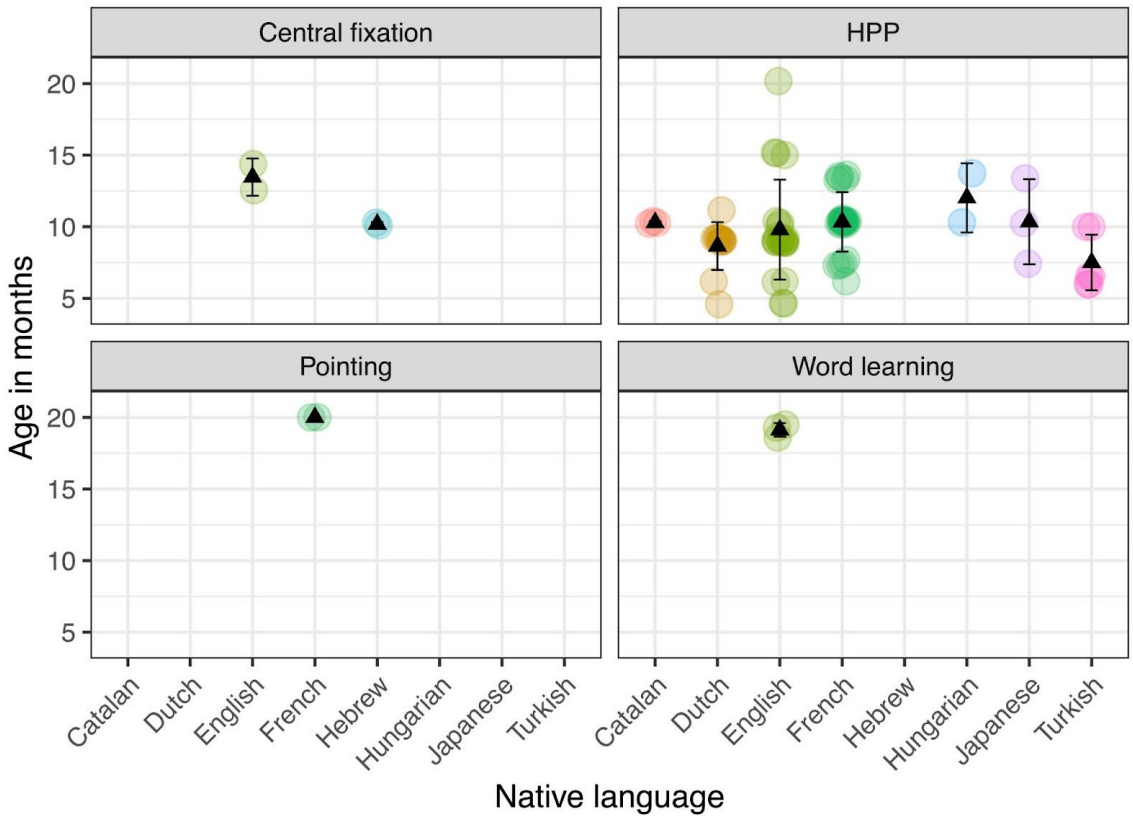

Figure 5. Infant age in months, split by native language and experimental method. Each colored point represents a study, the black triangle shows the mean for a given language, and error bars show one SD.

The distribution of infant age across methods and languages in our dataset is summarized in Figure 5. Examining the distribution of infants across methods it is clear that HPP was the experimental method of choice in almost all the experiments. With the caveat that the number of experiments varied greatly across methods, and that method was further confounded with age — for example, central fixation was used with younger infants, whereas pointing was only used with the oldest — we evaluated the moderating influence of the method on the effect size. We did this because variation in effect size as a function of method can inform experimental design for future research.

The estimates of the effect size along with the number of studies that used this method from a model with a fixed effect of Method (4 levels: *word learning*, *pointing*, *HPP*, *central fixation*) are listed in Table 1. Our results indicate that sensitivity to phonotactic patterns was least likely to be detected using pointing tasks (64%), with HPP, central fixation or even word learning being equally, and highly effective (>95%). Finally, the largest effect size was observed with central fixation. Note, however, that these estimates do not take into consideration the different ages of infants tested with each method. Therefore, they should be considered as guides only when running future experiments with infants of comparable ages to those included in the analysis here.

Test Method	Median ES	95% *CrI*	*p*	Experiments (*N*)
Word learning	0.62	[-0.06, 1.27]	0.96	3 (68)
Pointing	0.12	[-0.55, 0.77]	0.64	3 (58)
HPP	0.27	[0.09, 0.46]	0.99	74 (1,598)
Central fixation	0.75	[0.13, 1.37]	0.99	4 (74)

Table 1: Median effect size, 95% Credible Intervals and $p(\beta>0)$, with number of experiments included in the current meta-analysis for each of the methods used in the studies sampled here.

3.5 Infants are already tuned into phonotactics at 6-months

Next, we turned to the overall effect of age on sensitivity to phonotactics, the substantive question laid out in the introduction which motivated our study. Recall that sensitivity to native phonotactic patterns is thought to emerge only after 9-months (Jusczyk et al., 1994; and many others). In keeping with this developmental timeline first found by Jusczyk et al. in English-learning infants, the majority of infants tested cross-linguistically in subsequent years were also between 8- and 10-months (Figure 5). In fact, in summarizing our data, we found two natural breaks in the distribution of Age as shown in Figure 1; the first was around 8-months (235 days; see also Figure 1), and the second was around 10-months (325 days; Figure 1). We used these natural breaks to categorize age into three bins (*<8-months*; *between 8-10-months*; *>10-months*) in all analyses involving age in this paper. In keeping with conventional wisdom, we expected a significant positive effect size at the two older ages; further, if sensitivity to phonotactics develops between 6 and 9-months, we expected to find an aggregate effect size of 0 in infants below 8-months.

The estimates for the three levels of the Age predictor from the model described above are displayed in Table 2. Because the credible interval for infants below 8-months includes an effect size of 0, it is possible that they are not yet sensitive to phonotactic patterns. However, this outcome is quite unlikely. At all three ages the likelihood of a positive effect size was greater than 90%, with little difference between the age groups. The effect size at the two youngest age groups was also comparable. Thus, we found no evidence that sensitivity to native language phonotactic

patterns emerges between 6- and 9-months. Instead, the aggregate evidence from the meta analysis shows that infants are already sensitive to native language phonotactic patterns before 8-months.

Age Group	Median ES	95% *CrI*	*p*	Experiments (*N*)
< 8 months	0.25	[-0.10, 0.61]	0.92	17 (387)
8-10 months	0.29	[0.07, 0.52]	0.99	47 (1,044)
> 10 months	0.47	[0.13, 0.81]	0.99	20 (367)

Table 2: Median, 95% CrIs, and $p(\beta > 0)$ for effect size by age group, with the number of experiments included, and the total infants in that group.

3.6 Sensitivity varies by dependency type

Although in the previous section we presented an analysis that focused only on the infants' age, investigating age effects on the effect size is complicated by several factors. Most obviously, it is quite likely that *when* and to what extent infants tune into native patterns, varies from language to language.

This could be because languages differ in how much evidence supporting a specific pattern is available to infants. Consider the phonotactic pattern of vowel harmony. Vowel harmony is widely attested in many of the world's language families (for a review see Mintz, Walker, Welday & Kidd, 2018), where non-adjacent vowels in a word are constrained to be perceptually similar. Not all languages have vowel harmony — for instance, Hugarian and Turkish have harmony but not English — and even when two languages have vowel harmony, they may vary in the extent to which it is observed in the lexicon: for instance, vowel harmony is less pervasive in the Hungarian lexicon than in Turkish (see text accompanying results from Harrison, Thomforde,

& O'Keefe, 2004; see also Alderete & Finley, 2016). Thus, it is quite plausible that infants learning Turkish, because they have greater experience with vowel harmony, tune into it earlier than infants learning Hungarian.

Similarly, we know based on computational modeling that segment co-occurrence probabilities are more useful for finding words in English than in languages like French, Korean, or Japanese (Daland & Pierrehumbert, 2011; Daland & Zuraw, 2013; Boruta, Peperkamp, Crabbé & Dupoux, 2011). Because of differences in the usefulness of segment co-occurrence restrictions in their native language experience, English-learning infants tested on segment co-occurrence restrictions may demonstrate a larger effect size compared to infants learning French or Japanese.

There are also differences in the extent to which adult listeners rely on specific phonotactic patterns, even within typologically similar languages. English adult listeners' perception is affected to a greater extent by diphone probabilities than that of Dutch adult listeners (Smits, Warner, McQueen, & Cutler, 2003; Warner, Smits, McQueen, & Cutler, 2005; Park, Hoffman, Shin & Warner, 2018). It is possible that these differences in the cue weighting of co-occurrence restrictions across languages affect when infants learning these respective languages tune into them. Under such an account, English-learning infants would tune into diphone-based phonotactic patterns earlier or to a greater extent than Dutch-learning infants.

Another source of variation in effect size across languages could stem from qualitative differences among the patterns themselves. There is a large literature showing that it is harder for infants to learn non-adjacent compared to adjacent dependencies, whether they be over syllables or words (for a review see Wilson et al., 2020). Because vowel harmony involves tracking non-adjacent dependencies, infants may well be delayed in tuning into them compared to local segment co-occurrence restrictions.

Unfortunately, even with a dataset of almost 2000 infants learning 8 languages, the data were too unevenly distributed to evaluate the interaction of Age, Dependency Type and Language. Instead, given the unequal representation of data across languages (Figure 5), we evaluated the effect of the interaction between Age and Dependency Type, aggregated over Language. Differences between infants learning different Languages were probed only when sufficient data were available.

We considered three categories of phonotactic dependencies: non-local vowel dependencies (vowel harmony in Turkish and Hungarian; templatic melody in Hebrew), non-local consonant dependencies (in French and Japanese), and local dependencies (all others). Local dependencies, the largest subtype, included phonotactic patterns based on the frequency or positional frequency of a segment (unigram measures) and/or those based on the relative frequency of adjacent segments (bigram measures) as described in Jusczyk et al., (1994).

We separated the non-adjacent dependencies into those involving consonants and vowels for two reasons. Although infants' perception of vowel categories starts to become language specific by 6-months (see Tsuji & Cristia, 2014 for a meta-analytic review), their consonant categories do so only later in the first year (e.g., Tsao, Liu & Kuhl, 2006; Werker & Tees, 1984). Therefore, one might expect a similar asymmetry in the time course of infants' sensitivity to vowel dependencies relative to those defined over consonants. However, in experiments with artificial languages, infants have been reported to learn non-adjacent consonantal dependencies more easily than ones involving vowels (e.g., Bonatti, Peña, Nespor & Mehler, 2005, Newport & Aslin, 2004), so it is also possible that infants tune into non-local consonant dependencies earlier.

We first confirmed that the interaction between Age (*<8-months*; *between 8-10-months*; *>10-months*) and Dependency Type (*V-Nonlocal*, *C-Nonlocal*, and *Local*) was significant (β = -

0.81, 95% CrI [-1.48, -0.13], $p(\beta<0) = 0.99$). Given the significant interaction, we next analyzed the Age effects within each Dependency Type separately. The results are presented in Table 3, and summarized graphically in Figure 6. The results show distinct developmental trajectories based on phonotactic dependency type.

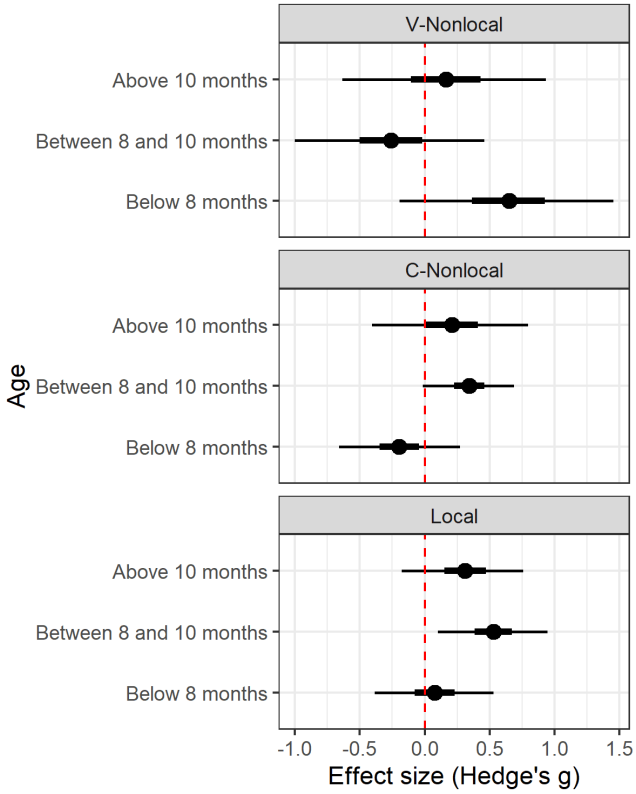

Figure 6: Effect size by binned Age crossed with Dependency Type, with the point representing the median, and the thicker and thinner intervals encompassing the central 50% and 95% Credible Intervals for values of effect size, respectively. The red dashed vertical line marks an effect size of zero.

Dependency Type	Age Group	Median ES	95% *CrI*	*p*	Experiments (*N*)

V-Nonlocal	< 8 months	0.64	[-0.20, 1.45]	0.93	4 (131)
	8-10 months	-0.26	[-1.00, 0.46]	0.23	7 (190)
	> 10 months	0.16	[-0.63, 0.93]	0.66	4 (60)
C-Nonlocal	< 8 months	-0.20	[-0.66, 0.27]	0.19	6 (104)
	8-10 months	0.34	[-0.02, 0.69]	0.97	13 (216)
	> 10 months	0.21	[-0.41, 0.79]	0.75	3 (44)
Local	< 8 months	0.08	[-0.39, 0.53]	0.63	7 (152)
	8-10 months	0.53	[0.10, 0.95]	0.99	27 (638)
	> 10 months	0.31	[-0.18, 0.76]	0.90	13 (263)

Table 3: Median and 95% CrIs and $p(\beta>0)$ for effect size by Age group nested within Dependency Type, with the number of experiments included, and the total infants in that group.

3.6.1 Sub-patterns in infants' sensitivity to non-local dependencies

For non-local vowel dependencies, there was a shift in the direction of preference around 8-months from familiar to novel. Infants younger than 8-months preferred high frequency sequences that were familiar, as indicated by the medium to large positive effect size; a positive effect size was also the most likely outcome of an experiment testing sensitivity at this age (93%). In contrast, infants between 8 and 10-months were more likely (~77%) to prefer less frequent sequences, i.e., those with novel non-local vowel dependencies, as indicated by the negative effect size. However, the evidence in support of a novelty preference was modest.

Interestingly, all experiments that contributed to the strong early sensitivity to non-local vowel dependencies investigated vowel harmony. From Figure 7 (left panel), we can see that infants learning a language with vowel harmony showed a large preference for harmonic sequences

before 8-months (β = 0.55, CrI = [-0.35, 1.36], p(β>0) = 0.89, based on 4 expts with 131 infants), with an almost equal swing towards novelty with increasing experience (β= -0.40, CrI = [-1.24, 0.49], p(β>0) = 0.18, based on 3 expts with 118 infants; i.e., 82% likelihood of negative effect size). Note that the results from infants older than 10-months were consistent with an effect size of 0, but unreliable because they were based on data from one experiment with only 14 infants (β = -0.13, CrI = [-1.06, 0.81], p(β>0) = 0.39). The findings on harmony are quite compelling because they are based on within lab comparisons of infants learning just two languages — Turkish and Hungarian — at different ages and using the same procedures.

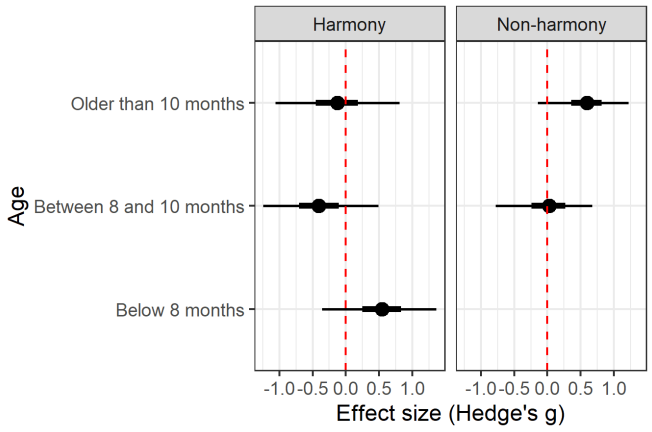

Figure 7. Effect size by binned Age crossed with type of nonlocal vowel dependency (harmony in the left panel, non-harmony in the right panel), with the point representing the median, and the thicker and thinner intervals encompassing the central 50% and 95% Credible Intervals for values of effect size, respectively. The red dashed vertical line marks an effect size of zero.

A different pattern emerged for non-local vowel dependencies that were not based on harmony (Fig 7, right panel). These data were from French- and Hebrew-learning infants' sensitivity to arbitrary co-occurrence restrictions between non-adjacent vowels. Only infants older

than 10-months preferred familiar native language sequences based on restrictions between non-adjacent vowels ($\beta = 0.59$, CrI = [-0.14, 1.22], $p(\beta>0) = 0.95$, based on 3 expts with 46 infants). No data were available for the youngest age group and the effect size in the 8- to 10-month-old bin was centered around 0 ($\beta = 0.03$, CrI = [-0.77, 0.67], $p(\beta>0) = 0.53$, based on 4 expts with 72 infants). This shows that infants tuned into non-local vowel dependencies other than harmony only after 10-months.

The developmental time course for non-local consonantal dependencies was somewhat similar to that for the non-local vowel dependencies that were not based on harmony. These findings were all on the dependency between labial and coronal consonants and almost exclusively from French-learning infants (French n = 344; Japanese n = 60). From Figure 6 and Table 3 we can see that not only did the credible interval for infants below 8-months include 0, the likelihood of a positive effect size was low, only 19%. In contrast, between 8 and 10-months, the effect size was always positive. Thus, our results confirmed that a sensitivity to non-adjacent consonantal dependency emerges between 8 and 10 months (Gonzalez-Gomez & Nazzi, 2012a; Nazzi, Bijeljac-Babic & Bertoncini, 2009; Gonzalez-Gomez, Hayashi, Tsuji, Mazuka & Nazzi, 2014).

In sum, infants tuned into non-adjacent consonant dependencies earlier than to non-adjacent vowel dependencies, but only when the latter were not harmony-based.

3.6.2 Infants are sensitive to local dependencies between 8- and 10-months

The last subtype, local dependencies, had the most diversity in terms of methodology, pattern tested, as well as native language background of infants. Collapsing across these categories, we confirmed that infants were sensitive to local dependencies in their native language both between 8-10 months and after 10-months (> 89%). Infants younger than 8-months, however, were

not sensitive to local dependencies; the effect size was centered near 0, with only a 62% likelihood of a positive effect size.

As can be seen from Figure 6, English (n = 614) and Dutch (n = 335) were the two languages with the most number of infants spanning different ages tested. In all these experiments infants were tested on local dependencies. This allowed us to investigate how language-specific differences might present within a specific subtype. The median effect size for each age group within the two languages is presented in Figure 8. The performance of infants learning English and Dutch differed at the two younger ages, but not when infants were older than 10-months. Infants below 8-months and between 8- and 10-months were more likely to demonstrate a positive effect if they were learning English (59% and 98%) not Dutch (37% and 87%). Even when the likelihood of a positive effect was overwhelming, that is between 8- and 10-months, the size of the effect was larger in infants learning English ($\beta = 0.60$, CrI = [0.05, 1.11], $p(\beta>0) = 0.98$, based on 15 expts with 344 infants) compared to Dutch ($\beta = 0.23$, CrI = [-0.21, 0.66], $p(\beta>0) = 0.87$, based on 9 expts with 222 infants). These differences in the likelihood of a positive effect as well as the magnitude of the effect size are consistent with a greater role of segment-segment co-occurrence restrictions in English phonotactics compared to Dutch.

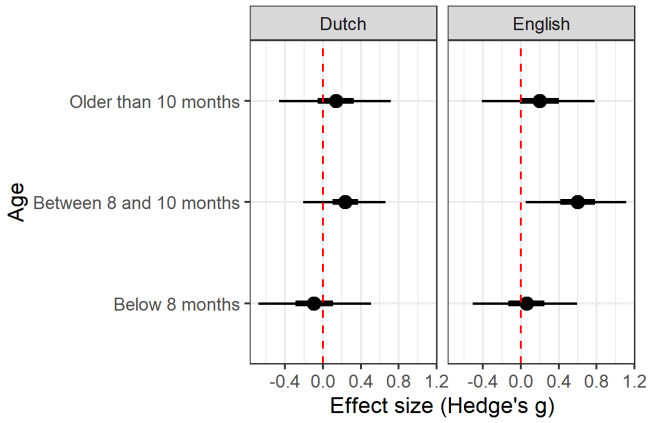

Figure 8. Effect size by binned Age for local dependencies in Dutch (left panel) and English (right panel). The point represents the median, and the thicker and thinner intervals encompassing the central 50% and 95% Credible Intervals for values of effect size, respectively. The red dashed vertical line marks an effect size of zero.

In summary, the results from the meta-analysis demonstrated that infants under 8-months were sensitive to phonotactics involving vowel harmony, but not other local or non-local dependencies. Between 8- and 10-months, infants demonstrated a novelty preference for vowel harmony and a robust familiarity preference for phonotactic patterns based on both local and non-local consonantal dependencies, with the former being larger. Only after 10-months was a consistent preference for non-local, non-harmony-based vowel dependencies evident. Finally, specific language experience altered the likelihood of a positive effect as well as its magnitude at the earliest stages of acquisition. In the next section we examined potential moderators that could affect the meta-analytic effect size for local dependencies, the subset of the data that was the largest and most heterogeneous.

3.6.3 Sensitivity is greater at edges, but unaffected by sonority sequencing violations

Finally, we evaluated two other moderators, one domain-general, and the other domain-specific, that have been proposed to influence the learning of local dependencies. First, we evaluated the influence of whether the pattern being tested was at the edge of a word. There is empirical as well as computational evidence that speech material at the edges of words, phrases or even sentences is very salient to infants (e.g., Daland & Pierrehumbert, 2011; Endress, Nespor & Mehler, 2009; Ferry, Fló, Brusini, Cattarossi, Macagno, Nespor & Mehler, 2016; Newport, 1990).

Speech material at edges is thought to be privileged based on a domain-general advantage in sensory encoding and recall (for a review see Hurlstone, Hitch & Baddeley, 2014; Sundara, 2018).

We only evaluated the edge effect within the set of local dependencies because all non-local dependencies involved segments at edges. Within local dependencies, the differences between effect size for phonotactic patterns at word edges and word medially was credible. Studies with phonotactic dependencies located word medially (10 studies with 176 infants) had an effect size of 0.29 (95% CrI = [-0.34, 0.90], $p(\beta>0) = 0.82$), whereas studies where the patterns were at an edge (37 studies with 877 infants) had an effect size of 0.46 (95% CrI = [-0.01, 0.89], $p(\beta>0) = 0.97$). The greater likelihood of the effect size as well as its larger magnitude confirmed that infants privilege speech material at edges for learning phonotactics. Figure 9 shows effect size as a function of edge.

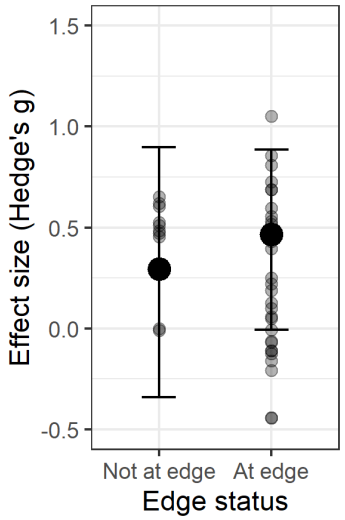

Figure 9. Effect size plotted as a function of edge status. Point estimates are the median of the posterior distribution for each group, with error bars encompassing the 95% Credible Intervals. Each translucent black dot represents an individual effect size from an experiment. Note that this plot excludes two outliers (X, Y, from condition "at edge") to make the differences more visible.

The final moderator we evaluated is more controversial. Based on typological tendencies, a language-independent linguistic bias referred to as the sonority sequencing principle is thought to govern the distribution of segments in specific syllable positions. Individual segments vary in sonority, typically depending on the extent to which the vocal track is open or the sound is loud (Parker 2002); thus vowels are considered to be the most sonorous while oral stop consonants (e.g. [p], [t]) are the least. Segment combinations that result in an increase in sonority from the onset of the syllable, and decrease in sonority at the offset of the syllable are the most common across languages. Although it is contentious whether knowledge of sonority sequencing is innate or learned from language experience, it has been shown to influence word segmentation as well as phonotactic judgments by adults (e.g., Daland, Hayes, White, Garellek, Davis, & Norrmann, 2011; Ettlinger, Finn, & Hudson Kam, 2012). Additionally, in one study, brain activity in newborns has been shown to differ in response to changes in sonority sequencing (Gómez, Berent, Benavides-Varela, Bion, Cattarossi, Nespor, Mehler, 2014). Thus, it is possible that infants' perception is influenced by sonority sequencing principles even in infancy.

This analysis was restricted to the subset of studies targeting CC clusters. If infants are sensitive to the SSP, we expected studies with low frequency items that violated them to have a larger effect size. However, a statistical analysis revealed no difference in the effect size of studies where low frequency items violated the SSP (4 studies with 87 infants; $\beta = 0.27$, 95% CrI = [-0.66, 1.19], $p(\beta>0) = 0.72$) compared to those that did not involve violations (1 study with 35 infants); $\beta = 0.20$, 95% CrI = [-0.76, 1.16], $p(\beta>0) = 0.66$). However, note that there was very little data for this comparison because in virtually all studies with CC clusters, low frequency items also violated sonority sequencing principles.

4 General Discussion

We aggregated data from 1924 infants learning 8 languages, tested using 4 methods across 84 experiments from 27 studies, to address when and how infants acquire native language phonotactics. We had three main findings from the meta-analysis. First, overall, infants favored the higher frequency phonotactic patterns as demonstrated by a positive effect size in the small-to-medium range. Second, infants were sensitive to some phonotactic patterns even before 9-months . Third, the developmental trajectory, as well as the size of the effect, differed substantially by dependency type: infants were sensitive to some non-local dependencies in their native language even before they were sensitive to local dependencies; this is even more unexpected because the non-local dependency involved vowels. We discuss each of these findings in turn.

We were motivated to conduct this meta-analysis by Cristia's (2018) findings that infants do not learn phonotactic patterns from short term exposure to an artificial language in the lab. This is surprising because with one exception, all studies summarized by Cristia (2018) tested patterns on word edges, a position that is privileged. In contrast with her findings, we found strong evidence of infants' ability to detect frequent phonotactic patterns from their native language. The natural and artificial language studies used similar methods and tested infants in a similar age range, so neither of these variables can account for the disparate findings. We suggest two alternative explanations.

The trivial explanation is that the extent of exposure in the artificial language experiments with infants, typically less than 4 minutes, is just too little for infants. In fact, it is not just phonotactics that infants fail to learn from short-term exposure to an artificial language; they also fail to learn phonetic categories (Cristia, 2018). Given evidence that sleep can facilitate generalization and consolidation of memory (see Gómez & Edgin, 2015 for a review), perhaps repeating short-duration exposure on consecutive days, or even testing on a second day, would

result in better outcomes. Alternatively, simply increasing the amount of exposure to around 20 minutes, as is typical in adult artificial language experiments that have been successfully replicated, may also serve the same purpose.

A second explanation may involve interference from the native language patterns that infants are learning at the same time. Outcomes of artificial language experiments have been reported to differ in adults learning different native languages attesting to interference effects (White et al., 2018). So it is possible that because of these interference effects infants were unable to learn from short term exposure in the lab. Regardless of the reasons, we can be sure that infants do not learn local dependencies, or even non-local consonant dependencies that are at the beginning and ends of words (hypothesized to be the easiest to learn; Endress & Mehler, 2010) from limited exposure in the lab. Whether infants can learn non-local vowel dependencies, particularly vowel harmony, from short term exposure to artificial languages in the lab, remains an open question.

One of the surprising findings of the meta-analysis was the early sensitivity infants demonstrated to vowel harmony. Infants demonstrated a significant, medium- to large-sized preference for harmonic sequences in their native language even before 8-months of age. This is unexpected because vowel harmony involves learning a non-local dependency and non-local dependencies are well-documented to be difficult to learn (for a review see Wilson et al., 2020). To account for the widespread prevalence of harmony across the world's languages, in some phonological frameworks like Autosegmental Phonology and Articulatory Phonology (Goldsmith, 1990; Browman & Goldstein, 1992), vowels are represented on a distinct tier, allowing V-V dependencies to be treated as local. Under such accounts, all V-V dependencies should be

privileged. However, this is not consistent with findings from infant experiments using either artificial or natural languages.

Research using artificial languages shows that infants find non-local dependencies between vowels to be particularly challenging to learn when they are arbitrary (e.g., Bonatti, Peña, Nespor & Mehler, 2005, Newport & Aslin, 2004 — neither included in the Cristia, 2018 meta-analysis). Indeed in this meta-analysis as well, we found evidence that infants tuned into non-local, non-harmony vowel dependencies only after 10-months. In contrast, infants tuned into non-local consonant dependencies between 8- and 10-months. So, infants tuned into non-adjacent, non-harmony dependencies involving vowels later in development than to ones involving consonants. In sum, not all V-V dependencies are privileged as might be expected if they are represented on a separate tier. Thus, the substance of the dependency itself, above and beyond its distribution in the stimulus items (local vs. non-adjacent) must also impact the developmental trajectory.

Not only were infants sensitive to vowel harmony at the earliest ages as demonstrated by a large familiarity preference, they showed a complete switch to a novelty preference between 8- and 10-months. A switch was not observed for any other dependency type, further attesting to the uniqueness of vowel harmony. A switch in preference for non-harmonic sequences after 8-months is likely to be a result of the unique experience infants learning vowel harmony languages have. Unlike other dependency types, languages with vowel harmony provide evidence for it in nearly every multisyllabic word. A switch to a novelty preference within a few months may be a result of this overwhelming experience with vowel harmony that infants have (e.g., Houston-Price & Nakai, 2004; Hunter & Ames, 1988).

It is also possible that it is easier for infants to learn vowel harmony compared to any other dependency type, even when provided with a comparable amount of evidence in their input,

because it is perceptually salient. Consistent with this explanation, there is some evidence that even English-learning infants who have no experience with harmony are able to use it to segment words (Mintz et al., 2018). If all infants are initially sensitive to vowel harmony, then our results show that sensitivity to vowel harmony for those with experience is facilitated before 8-months of age. Based on the findings of Mintz et al. (2018), it seems likely that for infants without experience with a native language exhibiting vowel harmony, the ability to detect vowel harmony declines at around the same age. Such a developmental trajectory is consistent with Attunement theories of perceptual development (Aslin & Pisoni, 1980; Aslin, Werker & Morgan, 2002). Attunement theories also provide the most comprehensive account of infants' discrimination of speech sound categories (Sundara, Ngon, Skoruppa, Feldman, Onario, Morgan, & Peperkamp, 2018).

The learning of all other phonotactic dependencies, including arbitrary non-local vowel restrictions, is induced by language experience consistent with Learning theories of perceptual development (Aslin & Pisoni, 1980; Aslin, Werker & Morgan, 2002). Our findings showed that the earliest patterns that infants acquire are likely to be local restrictions on the edges of words and phrases. Effect sizes were more likely to be positive, and larger for local dependencies at edges. In contrast, our results did not provide evidence that infants are sensitive to the sonority sequencing principle, although the data available were very limited. Whether this is because the effect size associated with the SSP is small, or because it varies to a large extent with language-specific experience remains to be determined.

At least for local dependencies, our results demonstrated that the timing of sensitivity, as well as the effect size of sensitivity to phonotactic patterns is affected by specific language experience. Infants learning English were more likely to show a positive, larger effect size for local phonotactic patterns compared to those learning Dutch learning. This was consistent with the

findings from adults that native listeners of English rely more on diphone probabilities than native listeners of Dutch. Further research is needed to tease apart the specific properties of the input infants are tuning into to extract local dependencies and how that input changes from infancy to adulthood.

By 8- to 10-months, infants are sensitive to both local and non-local consonantal dependencies. However, the effect size for infants' sensitivity to non-local dependencies remained smaller than that for local dependencies, consistent with proposals that the former are harder for infants to learn. A consistent, positive preference for non-local arbitrary vowel dependencies emerged even later — only after 10-months. This is also consistent with claims based on artificial languages that (arbitrary) non-local vowel dependencies are more challenging for infants to learn.

Infants older than 10-months were most likely to show a positive effect size for all dependencies — except for vowel harmony where the number of infants tested was very small. We judge that this effect is likely to be quite robust because infants in this age group were tested using all 4 methods and task demands in word learning or pointing experiments are quite different from those of listening preference experiments.

In sum, the results from this meta-analysis allowed us to generate evidence-based developmental trajectories for infants' sensitivity to different phonotactic patterns. Whether infants learn phonotactics from words in their lexicon (e.g., Thiessen & Saffran, 2003) or from the unsegmented speech stream (e.g. Adriaans & Kager, 2010; Brent & Cartwright, 1996; Daland & Pierrehumbert, 2011), we anticipate that the acquisition trajectory laid out in this paper will prove useful as a benchmark for constraining computational models of acquisition and for guiding future infant research.

5 Conclusion

In this meta-analysis, we aggregated data from around 2000 infants between 0 and 2 years of age, learning 8 languages, using 4 different methods from 84 experiments. Using Bayesian modeling, we established that unlike with artificial languages, in experiments with natural language stimuli, infants demonstrate sensitivity to phonotactic patterns. By 8-months, infants are already tuned into phonotactic patterns based on vowel harmony. Between 8- and 10-months, infants consolidate their learning of local restrictions as well as non-local restrictions involving consonants. Furthermore, infants' sensitivity to local restrictions is greater at word edges. Finally, sensitivity to non-local vowel restrictions that are not based on harmony are the last to emerge, and seen only after 10-months of age. By using Bayesian modeling in conjunction with the meta-analysis we were able to integrate the evidentiary value of nonsignificant and significant effects towards establishing the developmental timeline.

References

Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language, 62*(3), 311–331. https://doi.org/10.1016/j.jml.2009.11.007

Alderete, J., & Finley, S. (2016). Gradient vowel harmony in Oceanic. *Language & Linguistics*, *17(6),* 769–796.

Archer, S. L., & Curtin, S. L. (2011). Perceiving onset clusters in infancy. *Infant Behavior and Development*, *34*, 534–540.

Archer, S. L., & Curtin, S. (2016). Nine-month-olds use frequency of onset clusters to segment novel words. *Journal of Experimental Child Psychology*, *148*, 131-141.

Aslin, R. N., & Pisoni, D. B. (1980). Some developmental processes in speech perception. In G. H. Yeni-Komshian & J. Kavanagh & C. A. Ferguson (Eds.), *Child Phonology, 2: Perception* (pp. 67-96). New York: Academic Press.

Aslin, R. N., Werker, J. F., & Morgan, J. L. (2002). Innate phonetic boundaries revisited. *Journal of the Acoustical Society of America, 112,* 1257-1260.

Atlan, A., Kaya, U., & Hohenberger, A. (2016). Sensitivity of Turkish infants to vowel harmony in stem-suffix sequences: Preference shift from familiarity to novelty. In J. Scott, & D. Waughtal (Eds.), *Proceedings of the 40th Annual Boston University Conference on Language Development.* Online proceedings supplements.

Bergmann, C., & Cristia, A. (2016). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*, *19*(6), 901–917.

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, 1996-2009. DOI: 10.1111/cdev.13079 [Repository]

Black, A., & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 124-129.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science, 16*, 451-459.

Boruta, L., Peperkamp, S., Crabbé, B. & Dupoux, E. (2011). Testing the robustness of word segmentation with realistic input: effects of linguistic diversity and phonetic variation. *Proceedings of the 2011 Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*. Stroudsburg, PA: The Association for Computational Linguistics, 1-9.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1-2), 93-125.

Browman, C. P., & Goldstein L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3-4), 155-180.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1-28. doi: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. doi:10.32614/RJ-2018-017

Champely, S. (2021). pwr: Basic Functions for Power Analysis. R package version 1.3-0. https://CRAN.Rproject.org/package=pwr

Cristia, A., (2018). Can infants learn phonology in the lab? A meta-analytic answer. *Cognition, 170*, 312-327.

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology, 28*(2), 197-234. doi:10.1017/S0952675711000145

Daland, R., & Pierrehumbert, J.B. (2011), Learning Diphone-Based Segmentation. *Cognitive Science*, *35*, 119-155. https://doi.org/10.1111/j.1551-6709.2010.01160.x

Daland, R. & Zuraw, K. (2013). Does Korean defeat phonotactic word segmentation? *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), 873-877.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634.

Endress, A. D., & Mehler, J. (2010). Perceptual constraints in phonotactic learning. *Journal of Experimental Psychology: Human Perception and Performance, 36*, 235-250.

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, *13*(8), 348-353.

Ettlinger, M., Finn, A. S., & Hudson Kam, C. L. (2012). The effect of sonority on word segmentation: evidence for the use of a phonological universal. *Cognitive science*, *36*(4), 655–673. https://doi.org/10.1111/j.1551-6709.2011.01211.x

Ferry, A. L., Fló, A., Brusini, P., Cattarossi, L., Macagno, F., Nespor, M., & Mehler, J. (2016). On the edge of language acquisition: inherent constraints on encoding multisyllabic sequences in the neonate brain. *Developmental science, 19(3),* 488-503.

Friederici, A.D. & Wessels, J.M.I. (1993). Phonotactic knowledge of boundaries and its use in infant speech perception. *Perception & Psychophysics, 54*, 287.

Goldsmith, J. A. (1990). *Autosegmental and metrical phonology*. Basil Blackwell.

Gómez, D. M., Berent, I., Benavides-Varela, S., Bion, R. A. H., Cattarossi, L., Nespor, M., & Mehler, J. (2014). Language universals at birth. *Proceedings of the National Academy of Sciences*, *11*(16), 5837-5341.

Gómez, R. L., & Edgin, J. O. (2015). Sleep as a window into early neural development: Shifts in sleep-dependent learning effects across early childhood. *Child development perspectives*, *9*(3), 183–189. https://doi.org/10.1111/cdep.12130

Gonzalez-Gomez, N. (2012). *Acquisition of non-adjacent phonological dependencies: From speech percep-tion to lexical acquisition.* Université René Descartes - Paris V. PhD dissertation.

Gonzalez-Gomez, N, Hayashi A, Tsuji S, Mazuka R, & Nazzi T. (2014). The role of the input on the development of the LC bias: A crosslinguistic comparison. *Cognition*, 132, 301–311.

Gonzalez-Gomez N, & Nazzi T. (2012a). Acquisition of nonadjacent phonological dependencies in the native language during the first year of life. *Infancy*, *17*, 498–524.

Gonzalez-Gomez N, & Nazzi T. (2012b). Phonological feature constraints on the acquisition of phonological dependencies. In A. K., Biller, E. Y., Chung, A. E., Kimball (Eds.), *Proceedings of the 36th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press, 202–212.

Gonzalez-Gomez, N., & Nazzi T. (2012c). Phonotactic acquisition in healthy preterm infants. *Developmental Science, 15*, 885–894.

Gonzalez-Gomez, N., & Nazzi, T. (2013). Effects of prior phonotactic knowledge on infant word segmentation: the case of non-adjacent dependencies. *Journal of Speech, Language, and Hearing Research, 56*(3), 840-849.

Gonzalez-Gomez, N & Nazzi, T. (2015). Constraints on statistical computations at 10 months of age: The use of phonological features. *Developmental Science, 18*, 864-876.

Gonzalez-Gomez, N & Nazzi, T. (2016). Delayed acquisition of non-adjacent vocalic dependencies. *Journal of Child Language, 43*(1), 186-206.

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.

Gonzalez-Gomez, N., Schmandt, S., Fazekas, J., Nazzi, T., & Gervain, J. (2019). Infants' sensitivity to nonadjacent vowel dependencies: The case of vowel harmony in Hungarian. *Journal of Experimental Child Psychology, 178*, 170-183.

Graf Estes, K. (2014). Learning builds on learning: infants' use of native language sound patterns to learn words. *Journal of Experimental Child Psychology*, *126*, 313-27.

Graf Estes, K., & Bowen, S. (2013). Learning about sounds contributes to learning about words: effects of prosody and phonotactics on infant word learning. *Journal of Experimental Child Psychology*. *114*, 405-17.

Graf Estes, K., Edwards, J., & Saffran, J. R. (2011). Phonotactic constraints on infant word learning. *Infancy*, *16*, 180-197.

Harrer, M., Cuijpers, P., Furukawa, T. & Ebert, D. D. (2019). dmetar: Companion R Package For The Guide 'Doing Meta-Analysis in R'. R package version 0.0.9000. URL http://dmetar.protectlab.org/

Harrison, D., Thomforde, E., & O'Keefe, M. (2004). *The vowel harmony calculator.*. URL http://www.swarthmore.edu/SocSci/harmony/public_html/

Hohenberger, A., Altan, A., Kaya, U., Köksal-Tuncer, O., & Avcu, E. (2016) Sensitivity of Turkish Infants to Vowel Harmony: Preference Shift from Familiarity to Novelty. In N. Ketrez, & B. Haznedar (Eds.), *Studies in L1 Acquisition in Turkish*. John Benjamins Publishing, 22-56.

Hohenberger, A, Kaya, U, & Altan, A. (2017). Discrimination of vowel-harmonic vs vowel-disharmonic words by monolingual Turkish infants in the first year of life. In M, LaMendola, & J. Scott (Eds.), *Proceedings of the 41th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press, 309-322.

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development, 13*(4), 341-348.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. In C. Rovee-Collier, & L. P. Lipsitt (Eds.), *Advances in Infancy Research* (Vol. 5), Stamford: Ablex, 69-95.

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: an overview of the literature and directions for future research. *Psychological Bulletin, 140*(2), 339–373.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language, 32*(3), 402-420.

Jusczyk, P. W, Hohne, E. A., & Baumann, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics, 61*, 1465-1476.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*(5), 630-645.

Jusczyk, P. W. , Smolensky, P., & Allocco, T. (2002). How English-learning infants respond to markedness and faithfulness constraints. *Language Acquisition, 10*(1), 31-73.

Kabak, B. (2011). Turkish Vowel Harmony. In M. van Oostendorp, C.J. Ewen, E. Hume & K. Rice (Eds.), *The Blackwell Companion to Phonology,* 2831-2854, MA & Oxford: Wiley-Blackwell.

Kajikawa, S., Fais, L., Mugitani, R., Werker, J. F., & Amano, S. (2006). Cross-language sensitivity to phonotactic patterns in infants, *Journal of the Acoustical Society of America, 120*, 2278–2284.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. MA: Academic Press.

Lüdecke D (2019). *esc: Effect Size Computation for Meta Analysis (Version 0.5.1)*. doi: 10.5281/zenodo.1249218, https://CRAN.R-project.org/package=esc.

MacKenzie, H, Curtin, S, & Graham, S. A. (2012). 12-month-olds' phonotactic knowledge guides their word-object mappings. *Child Development, 83*, 1129-1136.

MacKenzie, H., Graham, S. A., & Curtin, S. (2011). 12-month-olds privilege words over other linguistic sounds in an associative learning task. *Developmental Science, 14*, 249-255.

Mattys, S. L., & Jusczyk, P. W., (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition, 78*(2), 91-121.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*(4), 465-494.

Mintz, T. H., Walker, R. L., Welday, A., & Kidd, C. (2018). Infants' sensitivity to vowel harmony and its role in segmenting speech. *Cognition*, *171*, 95–107. https://doi.org/10.1016/j.cognition.2017.10.020

Mugitani, R., Fais, L., Kajikawa, S., Werker, J. F., & Amano, S. (2007). Age-related changes in sensitivity to native phonotactics in Japanese infants, *Journal of the Acoustical Society of America, 122*, 1332–1335.

Nazzi, T. & Bertoncini, J. (2009). Phonetic specificity in early lexical acquisition: New evidence from consonants in coda positions. *Language and Speech, 52*(4), 463-480.

Nazzi T, Bertoncini J, & Bijeljac-Babic R. (2009). A perceptual equivalent of the labial-coronal effect in the first year of life. *Journal of the Acoustical Society of America, 126*, 1440-1446.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science, 14*, 11-28.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology 48*, 127-162.

Nicenboim, B., Roettger, T. B., & Vasishth, S. (2018). Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics, 70*, 39–55. DOI: https://doi.org/10.1016/j.wocn.2018.06.001

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*, 436–469.

Park, S., Hoffmann, M., Shin, P. Z., Warner, N. L. (2018). The role of segment probability in perception of speech sounds. *Journal of the Acoustical Society of America, 143*(3), 1920.

Parker, S. (2002). *Quantifying the sonority hierarchy*. University of Massachusetts Amherst PhD dissertation.

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/

Sebastián-Gallés, N., & Bosch, L. (2002). Building phonotactic knowledge in bilinguals: Role of early exposure. *Journal of Experimental Psychology: Human Perception and Performance, 28*(4), 974-989.

Segal, O., Keren-Portnoy, T., & Vihman, M. (2015). Infant Recognition of Hebrew Vocalic Word Patterns. *Infancy, 20*(2), 208-236.

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. *The handbook of research synthesis and meta-analysis, 2*, 257-277.

Smits, R., Warner, N., McQueen, J. M., & Cutler, A. (2003). Unfolding of phonetic information over time: A database of Dutch diphone perception. *Journal of the Acoustical Society of America, 113*(1), 563-574.

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of clinical epidemiology 54*, 1046-1055.

Sterne, J. A. C., & Harbord, R. M. (2004). Funnel plots in meta-analysis. *The Stata Journal*, *4*(2), 127-141.

Sundara, M. (2018). Why do children pay more attention to grammatical morphemes at the ends of sentences? *Journal of child language 45*(3), 1-14.

Sundara, M., Ngon, C., Skoruppa, K., Feldman, N. H., Onario, G. M., Morgan, J. L., & Peperkamp, S. (2018). Young infants' discrimination of subtle phonetic contrasts. *Cognition*, *178*, 57–66. doi.org/10.1016/j.cognition.2018.05.009

ter Haar, S. M., & Levelt, C. C. (2018). Disentangling Attention for Frequency and Phonological Markedness in 9- and 12-Month-Old Infants, *Language Learning and Development, 14*(4), 279-296.

Tsao, F.-M., Lui, H.-M., & Kuhl, P. K. (2006). Perception of native and nonnative affricate-fricative contrasts: cross language tests on adults and infants. *Journal of the Acoustical Society of America*, *120*(4), 2285-2294.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental psychology*, *39*(4), 706–716. https://doi.org/10.1037/0012-1649.39.4.706

Törkenczy, M. (2011). Hangarian Vowel Harmony. In M. van Oostendorp, C.J. Ewen, E. Hume & K. Rice (Eds.), *The Blackwell Companion to Phonology,* MA & Oxford: Wiley-Blackwell, 2963-2990.

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology, 56*, 179-191.

Van Kampen, A, Parmaksiz, G, Van De Vijver, R, & Hohle, B. (2008). Metrical and statistical cues for word segmentation: vowel harmony and word stress as cues to word boundaries by 6- and 9-month-old Turkish learners. In A. Gavarro, & M. J. Freitas (Eds.), *Language Acquisition and Development*. Newcastle Upon Tyne: Cambridge Scholars Publishing, 313–324.

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language, 103*, 151-175.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P., Paananen, T., Gelman, A. (2020). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. R package version 2.4.1, https://mc-stan.org/loo/

Warner, N., Smits, R., McQueen, J. M., & Cutler, A. (2005). Phonological and frequency effects on timing of speech perception: A database of Dutch diphone perception. *Speech Communication, 46*, 53-72.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49-63.

White, J., Kager, R., Linzen, T., Markopoulos, G., Martin, A., Nevins, A., Peperkamp, S., Polgárdi, K., Topintzi, N., & van de Vijver, R. (2018). *Preference for locality is affected by the prefix/suffix asymmetry: Evidence from artificial language learning.* In *Proceedings of the Forty-Eighth Annual Meeting of the North East Linguistic Society* (Volume 3). Graduate Linguistics Student Association, Amherst, MA, USA, 207-220. ISBN 978-1-72760-582-2

Wilson, B., Spierings, M., Ravignani, A., Mueller, J.L., Mintz, T.H., Wijnen, F., van der Kant, A., Smith, K. & Rey, A. (2020). Non-adjacent Dependency Learning in Humans and Other Animals. *Topics in Cognitive Science*, *12*, 843-858. https://doi.org/10.1111/tops.12381

Zamuner, T. S. (2006). Sensitivity to Word-Final Phonotactics in 9- to 16-Month-Old Infants, *Infancy, 10*(1), 77-95.

Appendix I

To re-analyze the 34 studies summarized in Cristia (2018) using Bayesian modeling, we used a Normal (0, 0.5) prior on the intercept, because of the finding of no significant directional effect size reported in Cristia (2018). This is in contrast to the model for natural language phonotactic sensitivity. Refitting the model with a Normal (1, 0.5) prior on the intercept did not change the results in a meaningful way. The intercept of this model gives an estimate of the group-level effect size.

Figure A provides a plot for the effect size for the studies from Cristia (2018), with the black dot and grey distribution representing the shrunken meta-analytic estimates of the true effect size for that study. Infants in this dataset range in age from 5-months to 17-months. And in all studies but one, infants were tested on local or non-adjacent consonantal dependencies at word edges. The vertical red dashed line marks an effect size of zero, and the grey vertical lines mark the median meta-analytic effect size, plus upper and lower bounds of the 95% CrI.

The effect size for this model based on the 34 experiments summarized by Cristia (2018) was 0.06, CrI=[-0.19, 0.32], $p(\beta > 0) = 0.68$. This is consistent with the results reported in Cristia (2018) that the effect size is zero, because the 95% CrI included 0. Further, the Bayesian method used here allowed us to more precisely characterize this null effect: a slight majority (68%) of the credible values for the effect size lie above zero consistent with a positive effect size, giving only the barest hint of the effect presumed in the literature. Even this small effect was, surprisingly, driven by experiments in which infants were taught non-local consonantal dependencies (Median effect size 0.17 [-0.35, 0.69], $p(\beta>0)=0.76$) instead of local ones (Median effect size = 0.01 [-0.36, 0.42], $p(\beta>0)=0.51$). Note that non-local dependencies are presumed to be harder to learn than local ones.

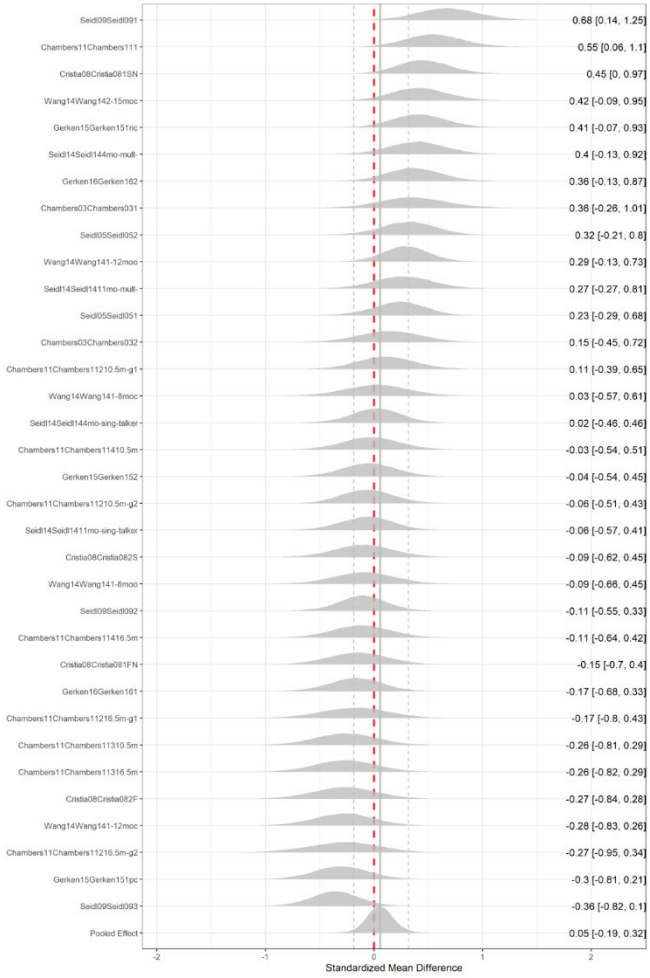

Figure A. Median and 95% Credible Interval for the shrunk effect size for each experiment summarized in Cristia (2018). The vertical dashed red line marks an effect size of zero, while the grey vertical lines mark the meta-analytic median effect size plus upper and lower 95% CrIs.