Quantifying Infants' Statistical Word Segmentation: A Meta-Analysis

Alexis Black (akblack2g@gmail.com)
Department of Linguistics, 2613 West Mall
Vancouver, BC, Canada V6T 1Z4

Christina Bergmann (chbergma@gmail.com)
Laboratoire des Sciences Cognitives, Ecole Normale Supérieure, 29, rue d'Ulm
75005 Paris, France

Abstract

Theories of language acquisition and perceptual learning increasingly rely on statistical learning mechanisms. The current meta-analysis aims to clarify the robustness of this capacity in infancy within the word segmentation literature. Our analysis reveals a significant, small effect size for conceptual replications of Saffran, Aslin, & Newport (1996), and a nonsignificant effect across all studies that incorporate transitional probabilities to segment words. In both conceptual replications and the broader literature, however, statistical learning is moderated by whether stimuli are naturally produced or synthesized. These findings invite deeper questions about the complex factors that influence statistical learning, and the role of statistical learning in language acquisition.

Keywords: language acquisition; statistical learning; word segmentation; meta-analysis

Introduction

Statistical learning (SL), the ability to extract statistical patterns from a continuous stream of perceptual experiences, is of fundamental theoretical importance. The first evidence that infants can extract statistical information from speech and use it to group syllables was provided by Saffran, Aslin, and Newport in 1996. This seminal paper has since accrued thousands of citations, and spurred a rich literature invoking SL as one foundation for language acquisition (see Newport, 2016) as well as perceptual learning more broadly (see Aslin, 2017). SL mechanisms have furthermore been successfully implemented in a range of computational models (e.g., Pearl, Goldwater, & Steyvers, 2010; Lloyd-Kelly, Gobet, & Lane, 2016). In short, statistical learning abilities are of fundamental, cross-disciplinary importance to better understand the computational foundations of cognition.

While many would accept some role for SL mechanisms, the nature and extent of this role remains contested. The more abstract the level of analysis, the more vigorous the debate (e.g., can SL yield syntactic ‘rules’?) – but inconsistencies emerge even at the level of tracking transitional probabilities (TPs) as a means of word segmentation. For example, the original effect has failed to replicate under certain conditions (e.g., variable word length: Johnson & Tyler, 2010; Lew-Williams & Saffran, 2012) or showed a developmental shift in cue-weighting (e.g., Thiessen & Saffran, 2003). Finally, a recent meta-analysis that examined natural speech word segmentation (not determined by TPs) revealed a significant, but small effect (Bergmann & Cristia, 2016), leading to concerns about the robustness of infants’ word segmentation in the absence of TPs.

In the current paper, we use meta-analysis to quantify and contextualize infants' ability to detect regularities in a continuous speech stream. To this end, we have aggregated all available evidence from the published record and present a meta-analysis of infant SL word segmentation studies. A meta-analytic approach helps establish the magnitude of an underlying effect, something single experiments are not equipped to do – and thus has the potential to impact future theory- and model-building. On the practical side, effect sizes are crucial for determining power of future studies, thus increasing the replicability of a line of inquiry and reducing the cost (failed studies, or testing too many participants) for single researchers.

We also take several steps beyond quantifying the underlying effect: Aggregating over studies allows for the identification of moderator variables, which also contributes to theory building and may guide future research. We examine three potential moderators that are relevant to the intersection of theories of infant cognition and statistical learning: (1) age, (2) stimulus naturalness, and (3) non-TP cues. The justification for investigating these particular moderators is described in brief. (1) All studies in the current meta-analysis use looking-time preferences. The direction of preference (to novel or familiar items) is commonly thought to relate to infant age and/or stimulus complexity (e.g., Hunter & Ames, 1988). We therefore predicted that developmental change might be reflected in a shift of preference (e.g., from a preference for words to one for non-words), or in a stronger effect over time. (2) Given the familiarity preference found in a previous meta-analysis on natural speech (Bergmann & Cristia, 2016), we hypothesized that it might be that the predominant novelty preference established for SL studies since Saffran et al. (1996) is grounded in methodological choices. The primary difference between these two datasets is in the nature of the stimuli: naturally produced vs highly artificial speech stimuli. Even within the literature of the current dataset, however, stimuli differ along this dimension. We therefore compare SL studies with natural and artificial stimuli. (3) Finally, a number of studies pitted alternative cues (e.g., word-level stress) against TPs. It is therefore important to examine the impact of these conflicting cues on SL performance compared to no conflict.

We also assess publication bias in the literature; a current topic that is especially important for infant research, considering the high cost of testing participants and the consequent use of small samples (Frank et al., 2017).

Methods

To collect data, we complemented expert lists with two google scholar searches. We first surveyed papers citing Saffran, Aslin, & Newport (1996) with the word "infant/infancy", but not "visual" in the title. The second search aimed to cast a wider net; search terms were now "month/s" and not "infant/infancy" or "visual". These two strategies yielded a total of 314 unique papers, which were then screened for inclusion. The criteria were: (1) contains data on infants from (2) behavioral experiments which exposed infants to a familiarization phase of continuous, artificial speech and which measured (3) reactions (typically looking times to unrelated visual stimuli) to both statistical words and non-words (this definition includes part-words).

The final sample encompassed 20 papers (10 containing conceptual replications[1]) yielding 68 (17 replication) effect sizes. Note that one paper often contains several experiments (henceforth: samples) that can yield effect sizes, for example when testing different age groups. In total, we are reporting on experiments testing 1,454 infants between 4.5 and 11.1 months. Children were tested in the headturn preference procedure (Kemler Nelson, et al., 1995; 59 samples) or the central fixation paradigm (Graf-Estes & Lew-Williams, 2015; 9 samples).

Effect Size Calculation

All scripts and raw data are available on github.[2] The effect size we report here is a standardized mean difference of infants' looking behavior when listening to statistical words versus non-words. Since a preference for non-words (novelty preference) is dominant in the literature, positive values reflect this direction of the effect. The larger the effect size, the bigger the observed standardized mean difference between the two types of test trials. In turn, negative values indicate that infants demonstrated a familiarity preference, i.e. they listened longer to statistical words over non-words[3].

We computed Hedges' *g* (Morris, 2010), a variant of Cohen's *d* (Cohen, 1988) that is preferred in the case of small sample sizes. Effect sizes were calculated based on reported test statistics: for 50 samples we could use means and standard deviations of test trials; for 17 samples *t*-values for the main comparison were available. To ensure consistency in the direction of the effect, we re-coded *t*-values as positive when infants listened longer to statistical non-words and as negative otherwise. We used standard formulae for effect size calculation in within-participant designs (Lipsey & Wilson, 2001, when means and standard deviations were available; Dunlap et al., 1996, for effect sizes based on *t*-values). One paper reported between-participant results and we computed effect sizes and variances from means and standard deviations accordingly (Lipsey & Wilson, 2001). When the same infants contributed to multiple effect sizes, we computed the median of all critical values to ensure independent samples (here, 4 effect sizes were derived from 8 non-independent samples). We could not compute effect sizes for 6 additional experiments, due to lack of information.

Only one of the 20 papers included reported correlations between test trials, which capture the dependency between the two data points stemming from the same participants and are necessary for *t*-value based effect size and general effect size variance calculation. We imputed random values based on the distribution of correlations reported in a similar meta-analysis (Bergmann & Cristia, 2016; updated data available via metalab.stanford.edu).[4]

Meta-Analysis

To establish the size and variance of the effect, we fitted a multivariate random effects model using the R (R core team, 2016) package metafor (Viechtbauer, 2010). Random effects models assume that all effect sizes are sampled from a distribution of effect sizes and try to estimate the mean and variance of this distribution. In the multivariate model, the interdependence between effect sizes from the same paper is taken into account, yielding a more robust measure

[1] A conceptual replication was defined as a study that did not introduce an additional, non TP-based cue, and did not differ from the original study protocol in a significant way. For example, studies that included a priming phase pre-familiarization, or a test phase involving carrier phrases were not included. See Github repository for a full list of included papers and the subset of conceptual replications.

[2] https://github.com/christinabergmann/StatLearnDB

[3] Given that infant looking-time studies generally accept either familiarity or novelty preferences, one might argue that we should instead use the absolute value of looking-time difference as dependent measure. Indeed, in the studies reported here that pit statistical learning against other cues, a switch in looking-time preference is explicitly predicted. We address these cues and their impact in the Complete Literature section of the paper. We would also like to address the general idea of absolute values in meta-analysis, and point to why this method may not be appropriate: 1) Theories of infant cognition and language acquisition have long sought to motivate the direction of looking-time preference; meta-analysis offers the potential power to test those theories and generate new possibilities when the theory is found to be inadequate. 2) Two opposing outcomes should reflect *two* underlying effects. Using raw effect sizes and testing the value of proposed moderators is a much more powerful use of meta-analytic techniques. Furthermore, it is important to recognize that allowing for two opposing outcomes, without the ability to predict those outcomes, increases the risk for false positives and might violate basic assumptions of sampling and null hypothesis significance testing.

[4] To assess the impact of this imputation, we re-ran our analysis with imputations based on varying means and verified that our conclusions about key findings do not change.

of the true effect. To investigate the impact of additional variables, we introduce moderators to this model.

Bias

We tested for bias in the published literature by assessing funnel plot asymmetry, which is significant when a portion of the expected distribution of effect sizes around the weighted mean is missing, yielding an over-representation of a part of the underlying effect size distribution. We test for asymmetry using the rank correlation test (implemented in metafor; Viechtbauer, 2010).

To further investigate biases, we make use of p-curves to test whether there is an excess in *p*-values just below the significance threshold of .05 and if the distribution of *p*-values indicates an underlying real effect (Simonsohn, Nelson, & Simmons, 2014). To this end, we enter all exact *t*-values that were reported (n = 48 for the whole dataset).

Results

Original Paper

We first calculated the effect size and its variance for the two experiments reported by Saffran, Aslin, and Newport (1996). Hedges' *g* was 0.4 (SE = 0.040) for experiment 1 and 0.38 (SE = 0.041) for experiment 2. According to Cohen's (1988) criteria this is a small to medium effect.

If experimenters base their sample size decisions on this effect size, they would have to test 53 infants in a paired samples design to achieve 80% power (computed with the R package pwr; Champely, 2016). The median sample size in our dataset is 22 participants, which would mean a 42% probability of obtaining a significant result, assuming the effect is of the size reported in the initial study; inversely, 58% of attempts to replicate this finding should fail.

Conceptual Replications

First, we report on the experiments that were identified as replications of the original report (Saffran et al., 1996). Seventeen experiments could be included in these analyses.

Meta-Analytic Effect The variance-weighted effect size Hedges' *g* is 0.21 (SE = 0.1), which is significantly different from zero (95% CI [0.02, 0.4], *p* = .03) and indicates a preference for statistical non-words. Note that this effect is smaller than the original report, and typical power is thus only 16% with 22 participants. Heterogeneity is significant, indicating variance in the data that is not explained by random measurement error (Q(16) = 71, *p* < .001).

Moderator Analysis: Age We find no significant effect of the moderator centered age in days (Q(1) = 0.6, β = -0.001, SE = 0.0015, 95% CI [-0.004, 0.018], *p* = .5).

Moderator Analysis: Stimuli Naturalness Studies on SL differ in the stimuli; in this dataset, 11 effect sizes came from experiments with synthetically generated speech, 6 were based on experiments with naturally produced speech.

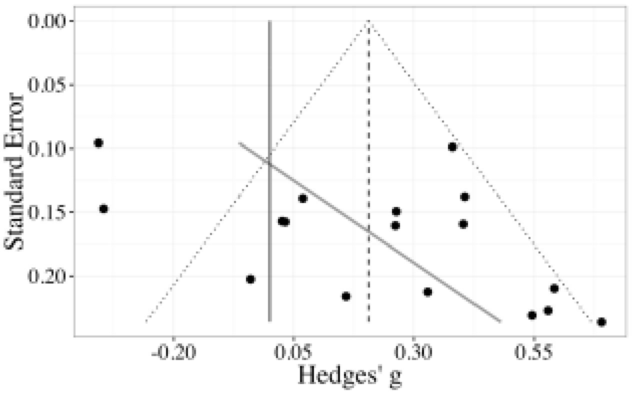

Figure 1: Funnel plot (code adapted from Sakaluk, 2016) showing standard error of the effect size as a function of effect size for 17 conceptual replications. The solid line marks zero, the dashed line the effect estimate, and the grey line indicates the funnel plot asymmetry.

Overall, the moderator test is significant (Q(1) = 5, *p* = .023) with a negative estimate (β = -0.35, SE = 0.16, 95% CI [-0.66, -0.05]), indicating that infants tend to show less of a novelty preference with stimuli produced by human speakers.

Follow-up analyses focusing on subsets revealed that synthetically produced stimuli lead to a significant positive effect (Hedges' *g* = 0.32, SE = 0.05, 95% CI [0.2, 0.4], *p* < .001), while those replications relying on naturally-produced speech yield an effect size not different from zero (Hedges' *g* = 0.02, SE = 0.2, 95% CI [-0.36, 0.41], *p* = .9).

Publication Bias The funnel plot shown in Figure 1 displays a greater density of large effect sizes that are of low-precision (lower right quadrant) and some effect sizes that are of high precision but outside the expected distribution (upper left quadrant), which is illustrated further by the linear regression line in grey. This line should be horizontal in the case of an even distribution around the median effect. Nonetheless, asymmetry is not significant with Kendall's τ = .26, *p* = .15.

The p-curve analysis based on the 6 significant *t*-values available in this dataset indicates a flat distribution of *p*-values, as would be expected when there is no underlying effect (Z = -0.43; *p* = .33). However, these 6 *t*-values might not be representative of the 17 studies analyzed here.

Complete Literature

Meta-Analytic Effect When taking into account all 68 independent effect sizes, the meta-analytic effect size Hedges' *g* is 0.09 (SE = 0.05), which is not significantly different from zero (CI [-0.02, 0.19], *p* = .1). This dataset, however, includes a number of samples that explicitly pit TPs against other segmentation cues, and thus may be expected to lead to different effects represented within the same data. Indeed, heterogeneity is significant (Q(67) = 334, *p* < .001). We thus analyze each of our moderators.

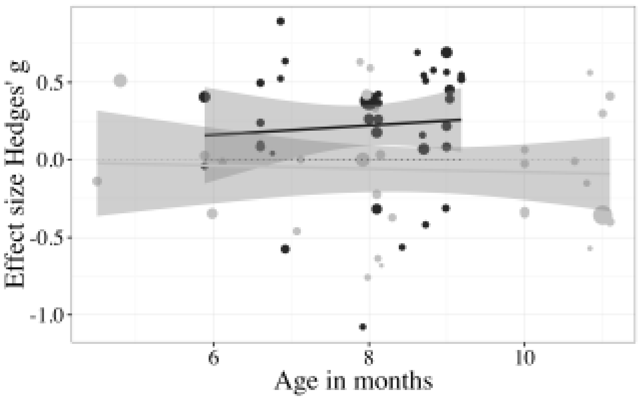

Figure 2: Effect size by participant age for all samples; point size is inverse variance. Black refers to synthetic, grey to natural speech. The dashed line indicates zero.

Moderator Analysis: Age As described in the introduction, more mature infants might show a different direction of preference or larger effect. However, we find no (linear) effect of age (Q(1) = 0.3, *p* = .6). Follow-up analyses introducing a quadratic term for age confirmed this finding.

Moderator Analysis: Stimuli Naturalness In the full dataset, the use of artificial and natural speech is fairly balanced, with 38 instances of computer-generated stimuli and 30 of human speakers. The moderator test is significant (Q(1) = 11, *p* < .001), and the results mirror our findings in the conceptual replication dataset. Figure 2 displays all samples, with color encoding natural (grey) vs artificial (black) stimuli. The meta-analytic effect for experiments with artificial stimuli is significantly above zero (Hedges' *g* = 0.23, SE = 0.06, 95% CI [0.11, 0.35], *p* < .001). In contrast, natural speech yields an effect not different from zero (Hedges' *g* = -0.05, 95% CI [-0.2, 0.06], *p* = .4).

Moderator Analysis: Cue conflict Cues can either be absent (n = 20), congruent with TPs (32), or in conflict with statistical information (16). Those cues encompass word stress (8), sentence level prosody (3), duration (2), intensity (2), and co-articulation (1). We predicted that cues that coincide with TPs might strengthen the effect, while those that conflict with TPs may reveal a different, possibly even opposing effect. We therefore introduced a three-leveled moderator. This analysis revealed no significant moderator effect (Q(2) = 1.9, *p* = .4).

Of the 48 samples that involve additional cues, 24 are based on the effect of a correlate of word-level stress on segmentation. These studies propose that infants will be driven to segment speech using a trochaic stress pattern, in line with their native language. Artificial languages with trochaic stress are therefore congruent with TP cues, and are predicted to lead (as a whole) to novelty preferences; those with iambic stress conflict with TPs, and are predicted to lead (as a whole) to null or familiarity preferences.

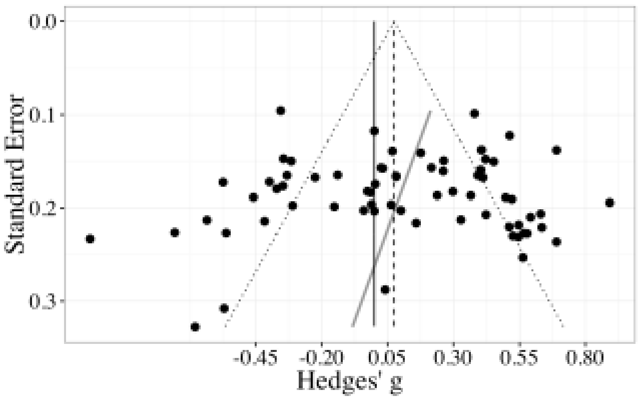

Figure 3: Funnel plot of all samples. For details see Figure 1.

A moderator analysis restricted to samples with additional stress-based segmentation cues fails to confirm this prediction (Q(1) = 0.7, *p* = .4; Cue conflict [iambic stress]: β = -0.07, SE = 0.08, 95% CI [-0.23, 0.09]).

Publication Bias Figure 3 shows an even distribution of effect sizes around the estimated median, the large spread illustrating the unexplained heterogeneity. The ranktest indicates no significant asymmetry (Kendall's τ = -.01, *p* = .9; see also grey linear regression line in Figure 3).

The p-curve based on 34 significant *t*-values indicates that the data contain evidential value (Z = -2.47; *p* = .007 for the full p-curve) and there is no excess of "just significant" p-values. Power based on the p-curve is estimated to be 25%.

Discussion

In the present paper, we examine infants' ability to track transitional probabilities (TPs) in continuous streams of speech. Experiments replicating the original Saffran et al. (1996) paradigm reveal a significant and reliable effect (Hedges' *g* = .21) that is on par with the effect found in the meta-analysis of natural speech segmentation (Hedges' *g* = .22; Bergmann & Cristia, 2016), albeit in the opposite direction of preference. An analysis of the whole literature fails to find a significant aggregated effect, but is reliably influenced by naturally vs. synthetically produced speech. There was no evidence for a developmental shift in or strengthening/weakening of preference, nor for a consistent and reliable role of additional cues. Finally, there is no clear evidence for publication bias. Taken together, these results invite deeper consideration of several issues in the future study of SL and theories of language acquisition, discussed in turns below.

One Mechanism Among Many

The data presented here confirm that infants can track statistically defined patterns and use that information to segment a stream of speech into word-like units. The strength of this capacity, however, may be more fragile than expected. How are we to understand these findings, as we continue to examine the import of statistical learning in language acquisition?

When aggregating across different studies, we put to the test the idea that researchers can predict the direction of infant looking-time preferences. Most popular theories of infant preference (e.g. Hunter & Ames, 1988; Kidd, Aslin & Piantadosi, 2012; 2014) predict an interplay between stimulus complexity and infant readiness to encode this complexity. In the case of TP-based word segmentation, we therefore expected a linear (or quadratic) shift from familiarity to novelty preferences as infants age. We instead find a consistent novelty preference. On the other hand, there is a significant effect of stimuli naturalness: While studies using synthesized speech yield reliable novelty preferences, studies using naturally produced speech fail to find reliable effects. It is likely that natural speech, even when altered to be largely monotonic and lacking syllable co-articulation, is more acoustically complex than synthetic speech. This is supported by the consistent familiarity preference across age groups found by Bergmann & Cristia (2016). Infants may thus be more likely to show a familiarity preference to natural speech because it may take more time to process (and hence habituate to/learn from) this complex signal. There is some evidence in the SL literature to support this idea: some studies find alternating patterns of looking preference by block (e.g. Graf-Estes & Lew-Williams, 2015). This, however, is rarely reported. Future investigations based on the meta-analytic data presented here might pursue the role of stimulus complexity by assessing the possible interactions between stimulus type, familiarization duration, age, and direction of looking-time preference.

Several of the studies in the dataset were designed to test the limits of SL. They have been included because in all cases infants might have opted to segment the language based on TPs alone; we hypothesized that, once taken in sum, these studies might have revealed evidence that TPs drive segmentation even in the face of alternative cues. This did not turn out to be the case – there is no reliable effect for segmentation when all studies are considered together. Moreover, and surprisingly, there is no pattern that unites samples in which cues are congruent with TPs vs those in conflict with TPs. These results, in fact, suggest that infants only succeed at tracking TPs when presented with artificial speech sounds. Given the results of the Bergmann & Cristia (2016) meta-analysis, we find this unlikely to reflect the true state of the world; rather, we believe it suggests that what does drive performance in the relatively simple paradigm of TP-based word segmentation remains underspecified and requires further theoretical, experimental, and meta-analytical consideration. Future work extending from the current dataset will aim to contribute to this discussion by accruing enough data to be able to examine additional moderators (e.g. familiarization duration) and outcome variables (i.e. effect sizes based on *proportions* of infants showing the effect, as opposed to standardized means of looking-time differences).

Practical Implications

There are several points to take into account when planning future SL word segmentation studies. First, assuming an effect size of Hedges' g = .21, the power of a typical 22 sample design is a meagre 16% (note that the p-curve analysis indicates an overall power level of 25% in the significant portion of the studies). A well-powered study (80%) would require a sample of 180 infants (142 if the direction of the preference can be predicted). This is impractical in the current state of infant research which relies on single labs conducting such studies (but see the alternative collaborative approach outlined by Frank et al., 2017). We do not intend to suggest that SL is not worth investigating – but it does call into question the methods with which we choose to investigate it. Power might, for example, be increased with more robust methods, calling for infant researchers to improve extant paradigms. At this point, we are only beginning to have sufficient power to fully understand the role of methods, stimuli, and test set-up (see e.g., Frank et al., 2017). One possibility lies in adopting more implicit measures of SL such as through neuroimaging, which may be less susceptible to factors affecting the direction of infant looking-preference.

Limitations

Any meta-analysis is limited by a number of factors, one of which is that the analysis is only as good as the data it contains. In other words, the studies reported here are those that have been published (or made available online) and were findable through our search criteria (see supplementary material for a full list of included studies). Since the effect is small, we expect that a number of failures to replicate the original finding are confined to the file-drawer, simply because they were underpowered. Further, studies showing a familiarity preference might not be published as those are not expected in replications of Saffran et al., (1996). Including such (presumed) file-drawer studies would make our estimates much more reliable and we strongly encourage researchers with unpublished work to contact the authors and contribute these findings (or any published data that may have been regrettably missed).

A second limitation is missing information. For example, in order to compute effect sizes and their variance for within-participant designs, it is necessary to know the correlation between infants' preferences for each test-item type. We have temporarily imputed these figures based on similar data (Bergmann & Cristia, 2016), and ran additional analyses to confirm that different values result in similar outcomes. However, we hope that authors who can retrieve this data will be willing to enrich our dataset, and recommend to all to include this information in future publications

Conclusion

This meta-analytic analysis of statistical learning as applied to word segmentation has revealed a reliable but small effect. We hope that this paper promotes future research that will seek to better characterize infant performance on SL tasks, and will thus contribute to stronger theories and models of infant cognition and behaviour.

Acknowledgments

The authors thank the Cognitive Science Society for hosting a meta-analysis tutorial in 2016 and for participants of this tutorial, who pre-coded some of the included papers. Alexis Black was supported by the UBC 4 Year Fellowship and NSERC Discovery Grant (Individual) to Carla L. Hudson Kam "Constraints on language acquisition and how they change (or don't) with age." Christina Bergmann was supported by the European Horizon 2020 programme (Marie Skłodowska-Curie grant No 660911), the Agence Nationale pour la Recherche (ANR-10- IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC), and the Fondation de France.

References

Aslin, R. N. (2017). Statistical learning: a powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*, e1373.

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

Champely, S. (2016). pwr: Basic Functions for Power Analysis. R package version 1.2-0. https://CRAN.R-project.org/package=pwr

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dunlap, W.P., Cortina, J.M., Vaslow, J.B., & Burke, M.J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170–177.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... Yurovsky, D. (under review). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. Preprint posted on *PsyArXiv* at https://osf.io/preprints/psyarxiv/27b43/

Graf Estes, K., & Lew-Williams, C. (2015). Listening through voices: Infant statistical word segmentation across multiple speakers. *Developmental Psychology*, *51*, 1517-1528.

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

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. Developmental science, 13(2), 339-345.

Kemler Nelson, D. G., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development*, *18*, 111-116.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS One*, *7*, e36399.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2014). The Goldilocks effect in infant auditory attention. *Child Development*, *85*, 1795-1804.

Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition*, *122*, 241-246.

Lewis, M., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P. E., Cristia, A., & Frank, M. C. (2017). A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis. Preprint posted on *PsyArXiv* at https://osf.io/preprints/psyarxiv/htsjm/

Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lloyd-Kelly, M., Gobet, F., & Lane, P. C. (2016). Under Pressure: How Time-Limited Cognition Explains Statistical Learning by 8-Month Old Infants. *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1476-1480). Austin, TX: Cognitive Science Society.

Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta- analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, *53*, 17-29.

Newport, E. L. (2016). Statistical language learning: Computational, maturational, and linguistic constraints. *Language and Cognition*, *8*, 447-461.

Pearl, L., Goldwater, S., & Steyvers, M. (2010). Online learning mechanisms for Bayesian models of word segmentation. *Research on Language and Computation*, *8*, 107-132.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

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

Sakaluk, J. (2016, February 16). 7. Make It Pretty: Forest and Funnel Plots for Meta-Analysis Using ggplot2. [Blog post]. Retrieved from https://sakaluk.wordpress.com/2016/02/16/7-make-it-pretty-plots-for-meta-analysis/

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.

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

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1-48.