Evaluating unsupervised word segmentation in adults: a meta-analysis

Joshua K. Hartshorne (joshua.hartshorne@bc.edu)
Department of Psychology and Neuroscience, Boston College
Chestnut Hill, MA 02467 USA

Wesley Ricketts (wesley.ricketts@bc.edu)
Department of Psychology and Neuroscience, Boston College
Chestnut Hill, MA 02467 USA

Abstract

Humans, even from infancy, are capable of unsupervised ("statistical") learning of linguistic information. However, it remains unclear which of the myriad algorithms for unsupervised learning captures human abilities. This matters because unsupervised learning algorithms vary greatly in how much can be learned how quickly. Thus, which algorithm(s) humans use may place a strong bound on how much of language can actually be learned in an unsupervised fashion. As a step towards more precisely characterizing human unsupervised learning capabilities, we quantitatively synthesize the literature on adult unsupervised ("statistical") word segmentation. Unfortunately, most confidence intervals were very large, and few moderators were found to be significant. These findings are consistent with prior work suggesting low power and precision in the literature. Constraining theory will require more, higher-powered studies.

Keywords: statistical learning; unsupervised learning; word segmentation; meta-analysis

Introduction

Humans, even from infancy, are known to be capable of learning many behaviors like language without receiving much direct feedback. The foundational work by Saffran, Newport, & Aslin (1996) provided initial evidence that infants can extract statistical regularities from auditory speech and leverage this information to identify words. This finding has precipitated the growth of a massive literature demonstrating that both infants and adults can engage in this behavior, known as unsupervised or "statistical" learning. Unsupervised learning is now widely considered to be an underpinning of human language acquisition. However, a number of important questions remain unanswered, including the extent to which unsupervised learning actually explains for language learning.

A central difficulty is that there are myriad learning algorithms for unsupervised learning, each with different assumptions and distinct implications (M. R. Brent & Cartwright, 1996; Chemla, Mintz, Bernal, & Christophe, 2009; Dupoux, 2018; Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Kurumada, Meylan, & Frank, 2013; Mareschal & French, 2017; Mintz, 2003; Monaghan & Christiansen, 2010; Perruchet & Tillmann, 2010; Swingley, 2005; Thiessen, 2017). These include algorithms that use tabulation of transitional probabilities, clustering algorithms, memory compression, recurrent neural networks, and inference over generative models —— each of which can be instantiated in a variety of manners (Frank et al., 2010, 2010; Newport & Aslin, 2004; Toro, Nespor, Mehler, & Bonatti, 2008).

While most of these algorithms can produce the basic findings — e.g., above-chance recognition of words subsequent to exposure to speech streams (Jenny R. Saffran et al., 1996; Jenny R. Saffran, Newport, & Aslin, 1996) — they have radically different implications. Different algorithms range widely in terms of how efficiently they learn and thus provide more or less plausible solutions for language learning (Batchelder, 2002; Michael R. Brent, 1999; Frank et al., 2010; Mareschal & French, 2017). They moreover suggest linguistic representations ranging from highly symbolic to essentially graded and distributed, and learning theories ranging from strongly empiricist to strongly nativist. Thus, determining which algorithm(s) humans actually use is not a minor technical point but rather is at the heart of the matter.

In sum, significantly constraining theory requires not only demonstrating that unsupervised learning takes place but also developing a precise quantitative understanding of how human statistical learning works, including what factors affect it and to what degree. This would allow precise comparison of human learning to different proposed algorithms.

Considerable effort has been made in this direction. Even just considering unsupervised auditory word segmentation in adults, well over 100 investigations have been published (see below); the numbers swell when one also considers investigations of other linguistic phenomena, learning in children, and visual statistical learning. However, using these results to constrain theory is not straightforward. There are a number of recent qualitative reviews of the literature (Armstrong, Frost, & Christiansen, 2017; Erickson & Thiessen, 2015; Frost, Armstrong, & Christiansen, 2019; Lidz & Gagliardi, 2015; Thiessen, 2017). However, qualitative reviews are of limited use in distinguishing theories that differ primarily quantitatively. Moreover, the literature contains a number of data-clashes, with different studies using slightly different methods and getting different results, or sometimes even using the same methods and getting different results (i.e., failures to replicate) (Bonatti, Pena, Nespor, & Mehler, 2005; e.g., Hartshorne et al., 2019; Newport & Aslin, 2004). While authors do try to adjudicate these data disputes, different researchers come to different conclusions (e.g., Bonatti, Pena, Nespor, & Mehler, 2007; Keidel, Jenison, Kluender, & Seidenberg, 2007).

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Table 1: Moderator Variables

Moderators	Levels	Description
Adjacency	Adjacent, Nonadjacent Syllables, Nonadjacent Consonants, Nonadjacent Vowels	Whether the relevant transition probabilities for word segmentation are between adjacent or nonadjacent segments.
Attention	No Task, Distractor Task, Overt Task	Task performed by the participant during the listening phase.
Average Syllables per Item	N/A	Avg number of syllables or tones per trained word.
Bilingual Extent	N/A	Extent participant speaks a second language.
Bilingual Immersion	N/A	Degree of immersion two languages
Foil Type	Nonword, Partword, TP-Match, Tone Foil	The design of the foils in the test phase.
Lengthening	True, False	Whether the initial/medial/final syllable of ea. word in the training is lengthened.
Length of Exposure	N/A	Duration of the training phase.
Musical Training	Yes, No	Whether the participant has musical training or not.
Native Language	English, Spanish, etc.	Native language of participants.
Phonotactic Match	True, False	Whether trained words conform to the phonotactic constraints of the participants' native language.
Pitch	Higher, Lower, False	Whether the initial/medial/final syllable of ea. word in the training has a modified pitch or F0.
Response Type	2AFC, Recognition	Response paradigm in the test phase.
Stim Type	Syllables, Tones	Type of stimuli composing the artificial language stream.
Target	Trained, Ruleword, Classword, TP-Match	Correct response in the test phase.
Total Number of Foils	N/A	Total num foils across all participants.
Number of Foils per Subject	N/A	Num unique foils each subject encountered.
Total Number of Trained Words	N/A	The total number of trained words present across all participants.
Number of Trained Words per Subject	N/A	The number of unique words a single subject encountered during training.

Overview of the Study

The goal of the present paper is to perform a quantitative summary of the literature — that is, a meta-analysis — to determine what quantitative constraints the literature places on unsupervised learning algorithms, if any. Meta-analyses provide a principled method for synthesizing divergent findings across experiments (Bailar, 1997; Borenstein, Hedges, Higgins, & Rothstein, 2009; Egger & Smith, 1998; Rothstein, Sutton, & Borenstein, 2006). There is one recent meta-analysis, though its scope was limited to effects of cue conflict and stimuli naturalness on unsupervised word segmentation in children ages 4 months to 11 months (Black & Bergmann, 2017). Morever, while these are theoretically important issues, they do not do much to distinguish learning algorithms.

We restricted our meta-analysis to unsupervised word segmentation in adults. We focused on adults because many more modulators have been tested for adults than for infants, giving us more power to potentially identify robust effects. With more than one hundred studies already published, we are also more likely to have sufficient power and precision to quantify effects than would be the case for the much smaller infant literature. While in principle the infant literature is more directly relevant to typical first-language acquisition, data to date provide little evidence of any age-related change in unsupervised word segmentation (Black & Bergmann, 2017; Raviv & Arnon, 2017; Thiessen, Girard, & Erickson, 2016).

The goal of the meta-analysis is to determine what moderators of statistical word segmentation are sufficiently well-evidenced to be strong constraints on theory – and, if possible, estimate effect sizes with enough precision to allow quantitative comparison to proposed algorithms.

Note that we leave actual comparison to algorithms for future work. This is partly because for many of these algorithms, it is not trivial to apply them to the kinds of challenges presented to humans in experimental settings, and thus determining their predictions is a significant endeavor in its own right (Frank et al., 2010). This is also because the results below suggest such comparison is premature.

Limitations

There are a few limitations of meta-analysis that bear mentioning. First, they are necessarily limited by the quality of the studies. Moreover, meta-analyses do not take a stand on the quality of any particular study, beyond anything captured by the standard error of measurement. With enough data, a few lower-quality studies should wash out, though of course this will not work if *most* studies have flaws in their methods.

Second and relatedly, meta-analysis necessarily elides differences between studies that are not captured by the covariates being considered. For instance, in our case, we did not note speech rate, which varies across studies and could conceivably affect results. To the extent that all studies used reasonable speech rates, this is actually beneficial, since the meta-analysis effectively captures uncertainty in measurement due to random variation in what is essentially a nuisance parameter. However, if there is theoretically-meaningful vari-

ation along some dimension that we did not code for, then in the best case scenario this adds undue noise, and in the worst-case scenario confounds the measurement of the moderators that were coded. Unfortunately, absent a perfect understanding of the phenomenon, deciding which moderators to code and which to ignore is a judgment call, and our judgment may not be correct.

The third and final limitation we wish to highlight is that given a fixed amount of data, the more moderators coded for in a meta-analysis, the less likely anything will be significant. This is particularly true if one corrects for multiple comparisons. Thus, there is always a tension between considering more moderators and thus avoiding the second problem listed above, and coding only the ones most likely to matter in order to preserve statistical power. One concession we made along these lines was to not consider interactions, with only a few exceptions. (Note that in some ways, this is a feature not a bug: the more information we wish to derive from the literature – that is, the more moderators we wish to measure – the more data there needs to be in the literature in order to have sufficient statistical power.)

Method

We based our methods on the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 checklist (Moher et al., 2015). Where our scientific priorities necessitated diverging from these criteria, we describe the differences and rationale below.

Eligibility Criteria

We created an inclusion/exclusion guide based on a preliminary survey of the literature. However, the sheer size and complexity of the statistical word segmentation literature are such that coders periodically found studies that did not clearly fit the criteria or which involved significant methodological variation that forced reevaluation of coding procedures. In these cases, the issue was resolved through discussion, the guide was updated, and the already-coded studies were double-checked for consistency. The final, exhaustive guide is available online (see OSF supplement). Ultimately, we selected studies meeting the following criteria:

Study designs The meta-analysis focused on auditory statistical word segmentation in adults. Participants must be familiarized with a continuous, artificial, auditory language stream and subsequently tested on their ability to explicitly discriminate between the words of this language and foils. In order to keep scope manageable and maximize comparability across effects, we excluded studies in non-auditory modalities; studies involving sentence-level phenomena such as artificial grammars or long-distance dependencies; and findings that did not involve binary choice or which involved implicit measures (e.g., neural responses or reaction times). The explicit measures of interest are most commonly instantiated in the form of a 2-alternative forced-choice paradigm or a yes/no recognition paradigm, where chance performance is at 50%. We make our exhaustive guide to inclusion criteria along with documentation of excluded studies available in our supplementary materials: osf.io/jmbdq/?view_only=5a9c63c532474a40b057abfeead1f119.

Participants Our sample is limited to empirical studies investigating healthy, adult humans in the general population (including college students ages 17 and older). Thus, studies conducted on younger populations or neurodivergent participants (e.g. those with dyslexia, ASD, brain lesions, etc.) were excluded. Studies comprising a mix of eligible and ineligible participants were only considered if the data from participants meeting our criteria were reported separately.

Information Sources

Due to resource limitations, we only included articles reported in the English language. Otherwise, our search protocol prioritized being exhaustive. This has the advantage of both including as much data as possible while simultaneously avoiding bias (since nearly everything is included). We began with a list of several hundred statistical learning studies that had been previously compiled, which we supplemented with suggestions from academic consultants, community listservs, and papers cited in two recent literature reviews (Frost et al., 2019; Jenny R. Saffran & Kirkham, 2018). As papers were marked for inclusion in the meta-analysis, we conducted targeted searches using key words suggested by the included papers, papers cited by included papers, and papers citing included papers. For instance, in July 2019, we used a Google Scholar search to find all papers published between 2015 and 2020 that cite Saffran et al. (-Jenny R. Saffran et al. (1996)) and contain the keywords "word segmentation," and then in December 2020 we performed a backward reference search of all included eligible reports that were published between 2015-2020. We continued this process until it stopped producing meaningful numbers of additional papers. This process yielded more than 700 unique articles that passed initial screening (the process described above does not easily yield itself to counting the total number of papers considered). In late 2021, we again used listservs and direct outreach to experts to identify any papers not already on our list; No new papers were included as a result.

Data management, selection, and collection

Each study identified from searching was screened independently by two or more members of the team. Data from eligible studies were manually extracted and entered into a plain-text spreadsheet in accordance with guidelines outlined in an instruction manual (available on OSF). The data recorded for each eligible study were verified by at least two members of the team, not counting periodic spot-checking and double-checking. Disagreements were addressed by discussion, often with additional members of the team, with all decisions documented. In order to promote consistency across reviewers, a small subset of eligible studies was used for training exercises prior to the start of the data entry process.

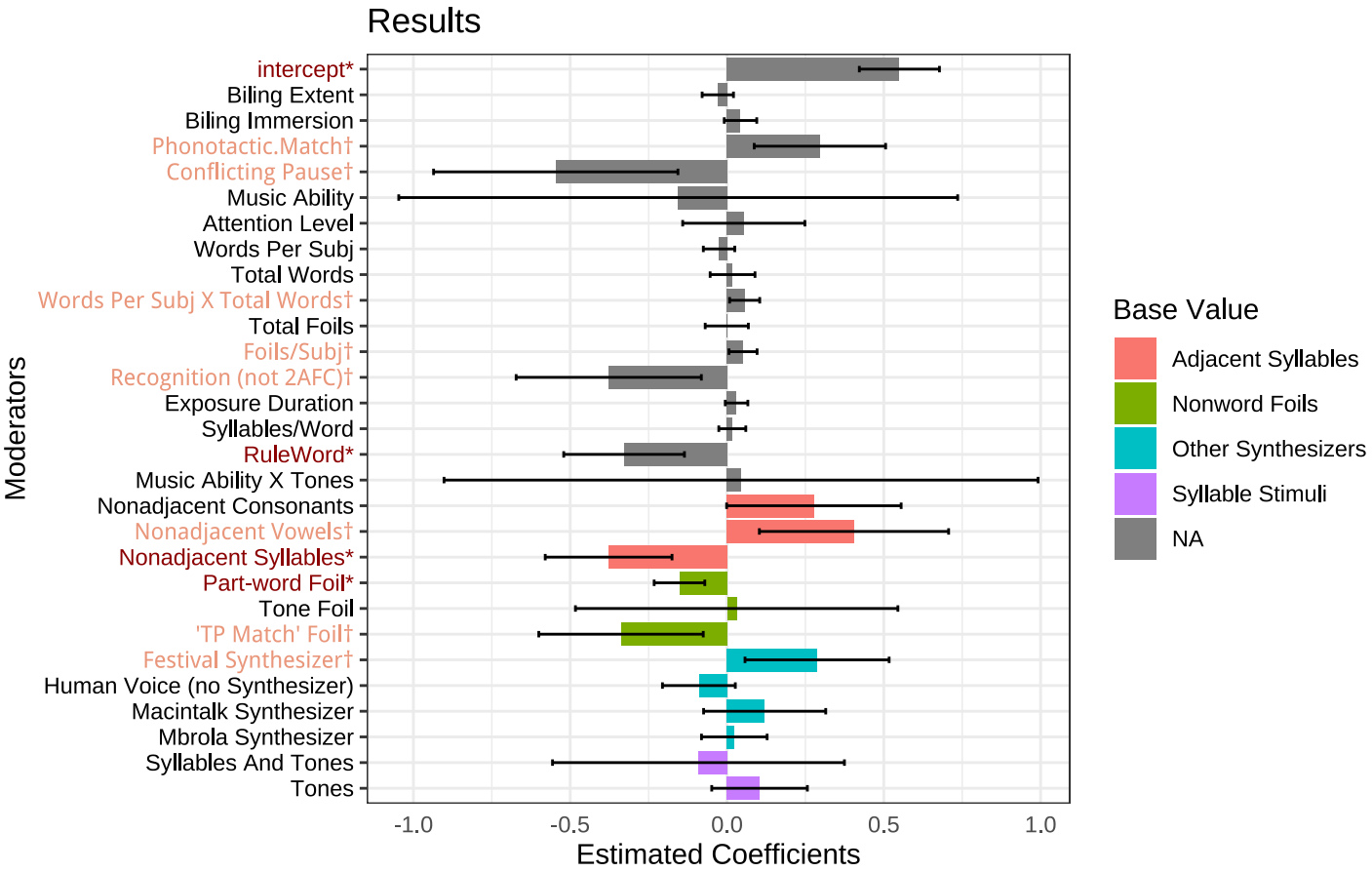

Figure 1: Results of meta-analysis including all moderators except native language and manipulations of lexical stress. * indicates significant after Šidák correction for multiple comparisons. † indicates significant without correction.

Exclusions were documented separately and consisted of (1) ineligible studies, including non-auditory statistical learning studies (N=933), and (2) studies with errors or which failed to report number of subjects, number of trials, or subject accuracy (N=17). Note that in some cases of studies with errors or missing information, we were able to obtain the critical information from the original authors; these studies are not included in the aforementioned counts.

Data Items

We coded single-sample effect sizes as the primary outcome measure. Ideally, these effect sizes would be extracted from mixed effects models that take into account both subject and item variability (Bates, Mächler, Bolker, & Walker, 2014). Unfortunately, the vast majority of published studies do not report such estimates, and many do not even report standard errors. Thus, we followed (Mahowald, James, Futrell, & Gibson, 2016) in terms of log odds assuming each data point is independent – a calculation that requires only knowing the mean response, the number of subjects, and the number of test items. (Note that this is possible since we are only including studies with dichotomous outcome variables.) While the assumption of independence is incorrect, (Mahowald et al., 2016) found that this estimate compared well to effect sizes extracted from mixed effects models (the gold standard). In situations where numerical data for participant accuracy was not provided in an experimental write-up, we used WebPlotDigitizer to manually extract this information from any available data visualizations (Rohatgi, 2021). The effect size and the variance for each experiment were then obtained via a generalized linear model in R.

Many studies report the same data analyzed different ways with different cell means. In a substantial portion of cases, more than one of these "slices" through the data are theoretically interesting (that is why the authors reported the data multiple ways). Rather than choosing one slice, we coded all of them (unless a particular slice ran afoul of exclusion criteria).

Selection of Modulators We compiled an initial list of modulators based on our preliminary survey of the literature, with a focus on modulators that appeared in at least 10 experiments. As the project continued, this list was revised as needed. A comprehensive guide was maintained and updated through (documented) discussion. Any changes prompted a review of already-coded studies. Table 1 contains a complete list of the modulators, along with the possible levels for each

(if applicable). For a more detailed description of each covariate, please refer to the coding manual in our supplementary materials.

Note that restricting the meta-analysis to modulators that are measured in a number of experiments excludes many that are highly theoretically-relevant. We view this as a limitation of the literature, not the meta-analysis. The goal was to identify well-evidenced modulators, the effects of which could be measured with some precision. By definition, this excludes effects that have only been investigated in a handful of experiments.

We ultimately excluded *mean age* from primary analyses both because it often could not be determined (21%) and because of the low variation (the vast majority of studies focused on college-age subjects). Other than native language (discussed below), other predictors had a maximum of 12% missingness with a median of 0%. In a very small number of cases where a level of a categorical variable was extremely rare (e.g., the "classwords" target type), those records were eliminated during analysis to prevent proliferation of covariates.

Data Synthesis and Analysis

Within a given experiment, the same data was often reported in multiple different ways reflecting different modulators of interest: for instance, comparing results for two conditions and also comparing results for bilinguals and monolinguals, collapsing across conditions. In order to capture both sets of results without double-counting data, we distributed the subjects across the slices: if an experiment had 50 subjects and 5 slices, we assigned 10 subjects to each slice. Note that in some cases where authors reported analyses that are not of theoretical interest, diluting the "key" analyses. Since reasonable people frequently disagree about which analyses are key – and because this estimation frequently changes over time – we did not adjudicate.

We conducted a two-stage meta-analysis using linear meta-regression as implemented in the *metafor* package in R (R Core Team, 2021; Viechtbauer, 2010). The first stage used all the predictors listed in Table 1 except for Native Language and two manipulations of lexical stress: Lengthening and Pitch, which were addressed in phase 2. We addressed incomplete records with multiple imputation as implemented in the *mice* package in R (van Buuren & Groothuis-Oudshoorn, 2011), averaging across 10 imputations and using a maximum of 50 iterations for convergence. After averaging across the imputations, we conducted the meta-regression with each moderator as a main effect and two theoretically-interesting interactions: between length of training and number of distinct words each subject must learn, and musical ability and whether the stimuli were pure tones.

In the second phase, we meta-regressed the residuals from the first phase against Native Language, the manipulations of lexical stress, and their interactions (Fig. 2). By addressing this analysis in a second phase, we avoided losing the 20% of records for which the subjects were bilingual or their native language was either unknown or variable across subjects – issues that were not correctable through multiple imputation. We considered the manipulations of lexical stress in this second phase because these manipulations are only interpretable in the context of the subjects' native languages.

All analyses were embedded in a reproducible manuscript using RMarkdown (Xie, 2018), which will be made available along with the other supplementary online material.

Results & Discussion

Our final sample comprised 367,821 unique observations from approximately 11,000 participants across 130 studies (note that many studies contain multiple experiments). Following APA guidelines for large meta-analyses, they are listed in online supplementary materials at https://osf.io/qp68h/?view_only=5a9c63c532474a40b057abfeead1f119.

Results for the first stage are shown in Fig. 1. The intercept reaches significance, reflecting overall above-chance word segmentation under our baseline conditions. That is, participants perform above chance when the words of the language are defined by high TPs between adjacent syllables, and when the foils presented in the test phase are "non-words" (i.e. strings that did not appear in the training). Three effects were found to significantly weaken performance: Subjects had more trouble learning when the "words" were defined by high TPs between non-adjacent syllables; when they were not tested on words observed during training but on "rule words" based on some pattern (typically same first and last syllable with a variable medial syllable); and when the trained words were pitted against "part-word" foils (which appeared in training) rather than "non-word" foils.

Results for the second stage – manipulations of lexical stress – are shown in Fig. 2. Only two effects were significant, with English-speaking subjects faring better when the last syllable of trained words dropped in pitch (relative to a monotone baseline) and French-speaking subjects faring better when the last syllable rose in pitch.

With the possible exception of the French finding, none of these results are unexpected. More unexpected was that not only were many moderators not significant, but this was rarely due to effects being estimated very close to zero. Rather, in most cases, the confidence intervals were very large, indicating a high degree of uncertainty (the non-effects of bilingualism and word-length being among the few counter-examples). The lack of statistical power across studies and the issue of multiple comparisons hinder the ability to investigate more complex interactions.

Conclusion

Despite well over 100 published papers on unsupervised word segmentation in adults, meta-analysis revealed few robust effects. In most cases, confidence intervals were quite large, suggesting a great deal of uncertainty about effect size. This is not entirely surprising, given prior evidence that unsu-

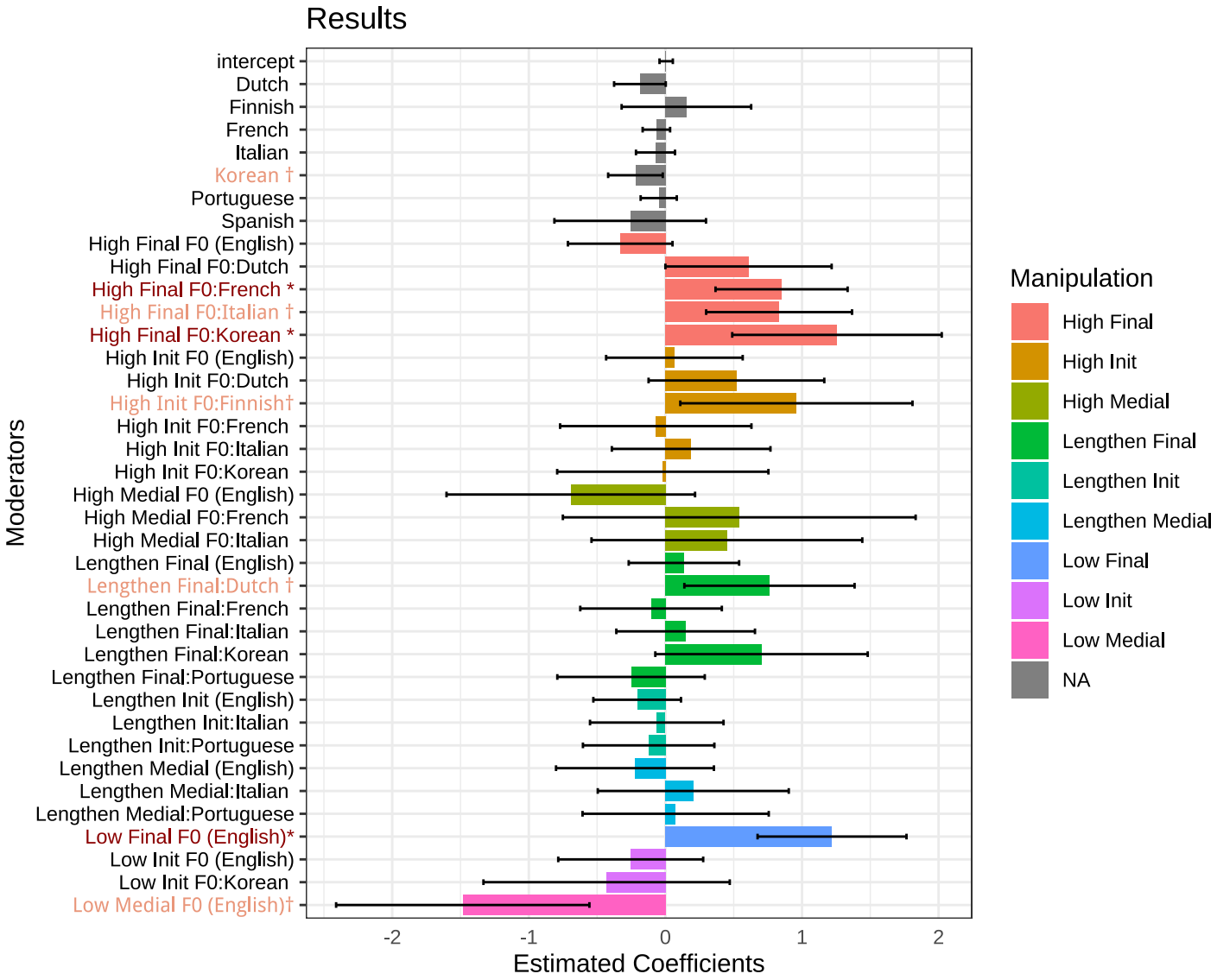

Figure 2: Meta-analysis of lexical stress effects, controlling for the moderators in Fig. 1. * indicates significant after Šidák correction for multiple comparisons. † indicates significant without correction.

pervised learning studies – like many literatures in the cognitive sciences – are substantially underpowered (Black & Bergmann, 2017; Collaboration et al., 2015; Hartshorne et al., 2019). In order to test theoretical proposals about what types of unsupervised learning humans are capable of, it will be necessary to conduct higher-powered studies.

Acknowledgements

This work was supported by grant 5108751 awarded by the National Science Foundation. We thank all authors whose studies were included and who provided raw data or clarification when requested. We also thank our board of academic consultants - Inbal Arnon, Joshua de Leeuw, Michael C. Frank, Hugh Rabagliati, and Mohinish Shukla for data and discussions. We are grateful to Lauren Skorb, Hayley Greenough, Tony Chen, Ning Duan, Rachel Duquette, Joyasha Johnson, Wendy Uelk, Everett Kim, Anna Petti, Alex Clendenning-Jimenez, and Alex Ichimura for their help with data collection and entry.

References

10 Armstrong, B. C., Frost, R., & Christiansen, M. H. (2017). The long road of statistical learning research: Past, present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*.

Bailar, J. C. (1997). The promise and problems of meta-analysis. *New England Journal of Medicine*, *337*(8), 559–561. http://doi.org/10.1056/NEJM199708213370810

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, *83*(2), 167–206.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014).

Fitting linear mixed-effects models using lme4. *arXiv Preprint arXiv:1406.5823*.

Black, A., & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. In *Proceedings of the 39th annual conference of the cognitive science society* (pp. 124–129).

Bonatti, L. L., Pena, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*(6), 451–459.

Bonatti, L. L., Pena, M., Nespor, M., & Mehler, J. (2007). On consonants, vowels, chickens, and eggs. *Psychological Science*, *18*(10), 924–925.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley; Sons. http://doi.org/10.1002/9780470743386

Brent, Michael R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*(1-3), 71–105.

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

Chemla, E., Mintz, T. H., Bernal, S., & Christophe, A. (2009). Categorizing words using 'frequent frames': what cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science*, *12*(3), 396–406.

Collaboration, O. S.others. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

Egger, M., & Smith, G. D. (1998). Meta-analysis bias in location and selection of studies. *BMJ*, *316*(7124), 61–66. http://doi.org/10.1136/bmj.316.7124.61

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, *37*, 66–108.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125. http://doi.org/10.1016/j.cognition.2010.07.005

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128.

Hartshorne, J. K., Skorb, L., Dietz, S. L., Garcia, C. R., Iozzo, G. L., Lamirato, K. E., … others. (2019). The meta-science of adult statistical word segmentation: Part 1. *Collabra: Psychology*, *5*(1).

Keidel, J. L., Jenison, R. L., Kluender, K. R., & Seidenberg, M. S. (2007). Does grammar constrain statistical learning. *Psychological Science*, *18*(10), 922.

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*(3), 439–453.

Lidz, J., & Gagliardi, A. (2015). How nature meets nurture: Universal grammar and statistical learning. *Annual Review of Linguistics*, *1*(1), 333–353.

Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, *91*, 5–27. http://doi.org/10.1016/j.jml.2016.03.009

Mareschal, D., & French, R. M. (2017). TRACX2: A connectionist autoencoder using graded chunks to model infant visual statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711).

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117. http://doi.org/10.1016/S0010-0277(03)00140-9

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., … Group, P.-P. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-p) 2015 statement. *Systematic Reviews*, *4*(1), 1. http://doi.org/10.1186/2046-4053-4-1

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, *37*(3), 545–564.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162.

Perruchet, P., & Tillmann, B. (2010). Exploiting multiple sources of information in learning an artificial language: Human data and modeling. *Cognitive Science*, *34*(2), 255–285.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raviv, L., & Arnon, I. (2017). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science*, *21*, 1–13.

Rohatgi, A. (2021). Webplotdigitizer: Version 4.5. Retrieved from https://automeris.io/WebPlotDigitizer

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. John Wiley & Sons.

Saffran, Jenny R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928. Retrieved from http://www.sciencemag.org/cgi/content/full/274/5294/1926?ijkey=4tT5xXrt3zjSo

Saffran, Jenny R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*(1), 181–203. http://doi.org/10.1146/annurev-psych-122216-011805

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

Swingley, D. (2005). Statistical clustering and the contents

of the infant vocabulary. *Cognitive Psychology*, *50*(1), 86–132.

Thiessen, E. D. (2017). What's statistical about learning? Insights from modeling statistical learning as a set of memory processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711).

Thiessen, E. D., Girard, S., & Erickson, L. C. (2016). Statistical learning and the critical period: How a continuous learning mechanism can give rise to discontinuous learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(4), 276–288.

Toro, J. M., Nespor, M., Mehler, J., & Bonatti, L. L. (2008). Finding words and rules in a speech stream: Functional differences betwen vowels and consonants. *Psychological Science*, *19*(2), 137–144.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. http://doi.org/10.18637/jss.v045.i03

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from https://doi.org/10.18637/jss.v036.i03

Xie, Y. (2018). Knitr: A comprehensive tool for reproducible research in r. In *Implementing reproducible research* (pp. 3–31). Chapman; Hall/CRC.