Predicting Early Language Development with Linguistic Alignment

Joseph Denby
Department of Psychology
University of Chicago
jgdenby@uchicago.edu

Daniel Yurovsky
Department of Psychology
University of Chicago
yurovsky@uchicago.edu

Abstract

Children are capable of quickly gaining enormous linguistic knowledge during early development, in part due to low-level features of their parents' speech. Some posit that parents contribute to their child's language development by calibrating their language according to their child's developmental abilities and needs. Here, we investigate this hypothesis by conducting a statistical analysis of this 'linguistic alignment' in a large-scale corpus of parent-child conversations recorded over a period of 5 years. Our results corroborate previous findings, showing strong parental alignment that slowly decreases as children mature; in addition, we demonstrate the impact of alignment on development by linking its effects with development outcome measures.

Keywords: Language acquisition; computational modeling; cognitive development

Introduction

Within their first few years of life, children make vast linguistic strides. Beginning in early toddlerhood, typically developing children gain over a thousand words in their vocabulary (as measured by the Communicative Development Inventories (Mayor & Plunkett, 2010) and quickly learn to produce unique, previously unheard utterances by combining known words in novel ways, becoming less repetitive over a short timespan (Lieven, Salomo, & Tomasello, 2009). Perhaps even more impressive, toddlers can infer dimensions of semantic content from words purely through exposure and differential use (Naigles, 1990; Yuan & Fisher, 2009).

To explain children's capacity for rapid language learning, some posit that children acquire a robust understanding of complex structures (such as language) by picking up on subtle statistical regularities present in their environment. Evidence for such a theory in language development spans childhood as well as linguistic features. For instance, infants can infer word segmentation in fluent speech, a particularly nettlesome early language task, through mere exposure (Saffran, Aslin, & Newport, 1996); with extended exposure, toddlers can extract syntactic information as well (Gomez & Gerken, 1999). Statistical learning undergirds vocabulary acquisition as well, with infants being capable of learning word-referent mappings through cross-situational exposure (Smith & Yu, 2008).

The aforementioned studies serve to flesh out the specific roles of low-level linguistic cues on early language learning. But, of course, most linguistic input doesn't come from a lab setting – it comes from real-world adults. Touched upon in previously discussed research, the linguistic tuning hypothesis puts forward the idea that parents bolster their child's early language learning by calibrating the complexity of their speech to the particular abilities and needs of their children (Montag & MacDonald, 2015; Snow, 1972; Thiessen, Hill, & Saffran, 2005). While the idea is intuitive and has some support, it faces some clear challenges. First, parents tend not to use simpler words when speaking to children, showing that there is no concrete tempering of vocabulary complexity for learning purposes (Hayes & Ahrens, 1988). Second, parents tend not to directly correct their child's syntactic errors, nor are children particularly receptive to corrections when they occur (Brown & Hanlon, 1970). While these findings point out a lack of overt calibration of speech to ability, there may still be subtle adjustments occurring (Sokolov, 1993; Spivey & Dale, 2006). For one, while overt syntactic correction may not occur (or be successful anyway), parents do appear liable to reformulate utterances that are ungrammatical, potentially providing 'negative evidence' for their child's grammatical benefit (Chouinard & Clark, 2003; Hirsh-Pasek, Treiman, & Schneiderman, 1984).

A parallel yet complementary vein of language development research investigates the presence of low-level cues in parental speech and their influence on child language learning. From this research, we know that child-directed speech contains features that facilitate language learning, and that more exposure tends to result in better outcomes (Weisleder & Fernald, 2013). Related, caregivers from families of high socioeconomic status (SES) tend to converse more with their children than their lower SES counterparts, and these increases are associated with improved development outcomes such as vocabulary size and school performance (Hoff, 2003; Walker, Greenwood, Hart, & Carta, 1994). So, given that more granular aspects of parent speech can have substantial effects on a child's language development, it may be that linguistic tuning occurs at this level in subtle ways, particularly when it comes to non-content words (i.e., words that are not central to the topic of discussion.)

This idea of assessing the direct impact of a parent's usage of non-content word on language development relates to linguistic alignment, a phenomenon whereby conversational partners tend to align aspects of their communicative style and content according to various external influences

e.g., (Pennebaker, Booth, Boyd, & Francis, 2015). Yurovsky, Doyle, & Frank (2016) investigates linguistic alignment in a natural language corpus of conversation between parents and children (CHILDES) (MacWhinney, 2000) to assess whether tuning occurs at the level of syntactic categories. Through this analysis, they find that syntactic alignment does indeed exist in both parents and children; moreover, parents exhibit a negative alignment trajectory over time, suggesting that the relationship their speech shares with their child's changes as a function of development. These results present a powerful proof of concept that syntactic alignment exists between parents and children, but it remains unclear whether alignment bears any sort of concrete relationship to language development.

Here, we extend their model by applying it to the Language Development Project (LDP) (Goldin-Meadow et al., 2014), a corpus of ecological conversations between parents and their children over time, collected from a socioeconomically diverse sample of parent-child dyads. Moreover, we use alignment estimates alongside demographic information to predict language outcome measures, strengthening the linguistic tuning hypothesis by concretely showing how parents' sensitivity to their child's linguistic needs and abilities impacts their development.

Model

The linguistic tuning hypothesis predicts that parents will calibrate their language in part by assessing their child's needs and abilities. So, we predict that parents will exhibit high alignment to their young children, but will reduce their alignment as their children mature. To test this prediction, we employ an extended version of the Hierarchical Alignment Model (HAM) implemented in Yurovsky et al. (2016) which both estimates the impact of a speaker's use of syntactic categories on their conversational partner's usage and uses these alignment estimates to predict language outcome scores.

At base, the model predicts, for each utterance, whether the speaker will produce a word from a given syntactic category. This prediction is generated by two factors: the speaker's baseline propensity towards using that category and the speaker's tendency to align, producing words from a category just used by their partner. In the model, the primary computation mimics a standard logistic regression - the production of a syntactic category within an utterance is treated as a binary outcome variable impacted by a linear combination of predictor variables (here, baseline usage and alignment.) The model's hierarchical structure then allows the estimates of baseline usage and alignment effects to be pooled across individual speakers and syntactic categories in a way that ensures statistical robustness.

The model used here then incorporates these alignment and baseline usage estimates as predictors in a linear regression model of PPVT (Dunn & Dunn, n.d.). At this stage, PPVT is estimated as a linear combination of predictors reflecting alignment and baseline usage estimate for both parents and children alongside other predictors representing demographic variables (e.g., child's gender, mother's education) and the child's age.

Model Details

The structure of the model used here greatly resembles that used in Yurovsky et al. (2016), in that it operates over utterances represented as binary vectors, with indices indicating the presence or absence of each of the 14 LIWC categories. The probability of producing each category in each utterance is predicted from two parameters: the speaker's baseline usage of that syntactic category (η^{base}), and the change in that speaker's baseline as a function of interacting with the listener (η^{align}). So, for replies to utterances that don't contain a given syntactic category, the production parameter for that category is computed by applying the inverse logit function to the appropriate baseline log odds:

$$P(Production) = logit^{-1}(\eta^{base})$$

Alternatively, replies to utterances that *do* contain a given syntactic category, the parameter computation takes into account the sum of the baseline and alignment log odds:

$$P(Production) = logit^{-1}(\eta^{base} + \eta^{align})$$

To accommodate the variance in production across the LIWC categories, each baseline usage parameter was drawn from an uninformative prior ($\eta^{base} \sim Uniform(-5,5)$); alignment parameters were regularized towards 0 by way of implementing a conservative prior ($\eta^{align} \sim Normal(0,1)$).

All parameters were estimated hierarchically, which allows intelligent pooling of data across participants in the dataset. Each subpopulation (i.e., all parents and all children) obtained an alignment estimate, each of which produced speaker-level alignment estimates, which produced category-level alignment estimates. The order was flipped for baseline parameters in order to better reflect empirical baseline usages across syntactic categories; subpopulation estimates produced category-level estimates, which then produced speaker-level estimates. As in Yurovsky et al. (2016), we also include parameters that allow baseline (β) and alignment probabilities (α) to change linearly over time.

Next, we extend the model used in Yurovsky et al. (2016) by using estimated parameters to predict PPVT scores, a widely used inventory for tracking language development. To do so, we effectively implemented a linear regression model, where PPVT scores were modeled as linear combinations of various predictor variables. These predictor variables included the child's age (κ^{age}), alignment parameter estimates for the child and their parent ($\kappa_c^{align}, \kappa_p^{align}$), the mother's education (κ^{ed}), the child's gender (κ^{female}), as well as interaction effects for all variables with age. The error variance for this model was estimated using parameter σ.

The model implemented here then serves two purposes: (1) It extends the analysis of Yurovsky et al. (2016) to a new

dataset, aiming to replicate previous findings in a more diverse and representative sample, and (2) It incorporates alignment estimates in a predictive model of early language outcomes, serving to test the hypothesis that alignment has significant effects on language development over and above demographic features. To be specific, we hope to replicate non-zero estimates for η parameters (demonstrating that alignment between parents and children exists across datasets), positive β for children (showing that children increase their baseline usage of categories over time), and negative α for parents (showing that parents decrease their alignment as their children age.) Secondly, if the PPVT model estimates for parameters corresponding to the main or interaction effects of alignment are non-zero in the presence of demographic variables, we can infer that alignment has a significant effect on early language development.

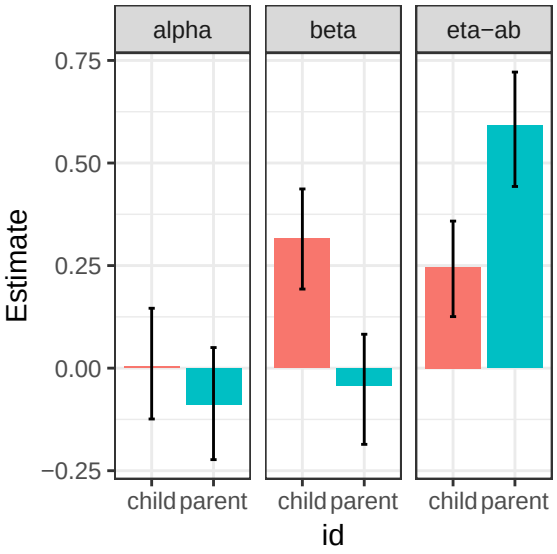

Figure 1: This plot is not beautiful yet.

Analysis

Data and Methodology

Transcripts were sourced from the Language Development Project corpus (Goldin-Meadow et al., 2014), consisting of conversations between 62 child-parent pairs spanning 12 to 60 months of age. This corpus amounts to 712 total transcripts with a median representation of 12 sessions per subject pair. The sample includes 30 female children and large representation of socioeconomic status based on the mother's level of education: 22 with an advanced degree, 20 with a bachelor's degree, 10 with some college or trade school, 8 with a high school degree or GED, and 2 with some high school.

Following Yurovsky et al. (2016), successive utterances from a speaker within a transcript were concatenated into a single utterance. Individual utterances were then transformed into a binary vector with indices indicating the presence or absence of each of the 14 LIWC categories. This pre-processing ensured turned every transcript into a speaker-reply format: each utterance within a transcript was both a reply to the preceding utterance and a message to the next one.

Each transcript was then compressed, yielding 4 numbers for each LIWC category. For a pair of speakers A and B in a transcript, for each LIWC category, we computed the number of utterances from A to B containing the category (N^{align}), the number of utterances from A to B not containing the category (N^{base}), the number of utterances containing the category responding to an utterance containing the category (C^{align}), and the number of utterances containing the category responding to an utterance not containing the category (C^{base}). Aggregating in this way provided the platform for the model's sampling - for each transcript, C^{base} and C^{align} were drawn from Binomial distributions parameterized by N^{base} and N^{align} chances respectively, with probabilities computed via the logistic regression models outlined above.
Sampling was performed using Stan, a probabilistic programming language that implements Hamiltonian Monte Carlo sampling methods (Carpenter et al., 2017). Posterior distributions for each parameter in the model were estimated using 350 iterations (THIS WILL LIKELY CHANGE). [1]

Results

Alignment estimates for parents and children were both significantly above zero, corroborating the findings of Yurovsky et al. (2016) in showing that both groups exhibit alignment (Figure 1). We also replicate the finding that parents appear to align more to their children than children align to their parents. The model estimates developmental baseline changes at approximately zero for parents, but significantly above zero for children, replicating previous findings. Alignment is estimated as having no significant age effect in this dataset, failing to replicate the early finding that parents tend to exhibit decreased alignment over their child's development. However, the mean estimate is trending negatively, suggesting this may be a function of the data's limited scope (Figure 2).

The estimates for PPVT predictors are listed alongside their standard errors in Table 1. From the estimates we see, as we might expect, that PPVT is positively associated with the age of the child, the level of education of their mother, and their being female. Moreover, female children tend to have a decreased age effect on PPVT. Alongside these demographic effects, we see some robust alignment effects on PPVT: parental alignment is associated with increased PPVT, but a decreased age effect, while child alignment is associated with decreased PPVT and an increased age effect.

[1] Code available at `https://github.com/callab/ldp-alignment`

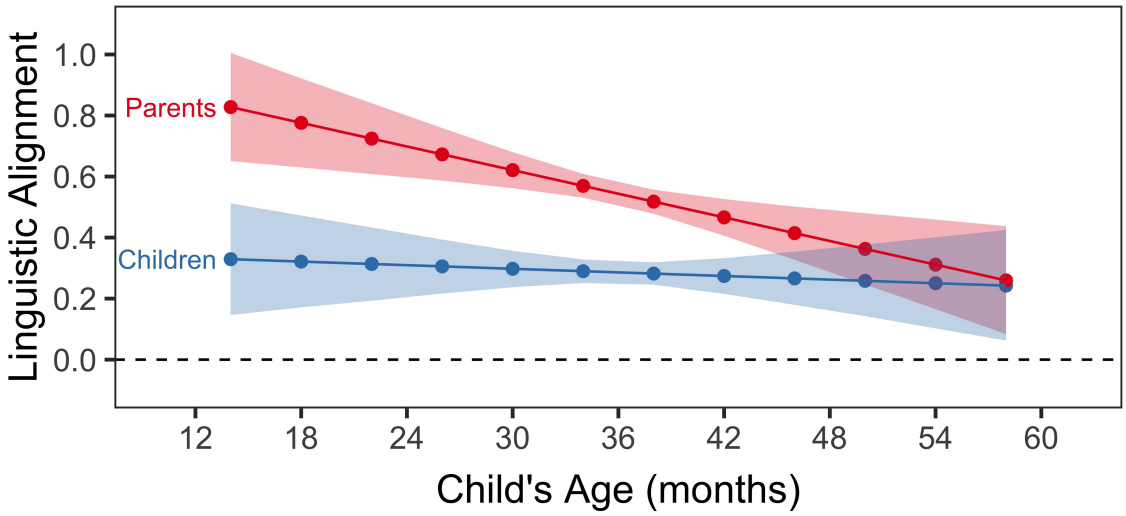

Figure 2: This is not the right plot but it is close.

	measure	Estimate	StandardError
1	ppvt_age_years	18.16	1.59
2	ppvt_child_align	-100.08	14.82
3	ppvt_child_align_slope	37.09	3.99
4	ppvt_education	5.37	1.62
5	ppvt_female	16.48	3.98
6	ppvt_female_slope	-1.90	1.29
7	ppvt_intercept	-65.31	4.93
8	ppvt_mother_ed_slope	-0.01	0.52
9	ppvt_parent_align	62.66	2.83
10	ppvt_parent_align_slope	-6.33	1.14
11	sigma	4.57	0.01

Table 1: Parameter Estimates for PPVT predictors with standard errors.

Discussion

In an effort to understand and investigate how children rapidly acquire language, some have put forward the linguistic tuning hypothesis, arguing that the language children hear is somehow calibrated to their particular needs and abilities (Snow, 1972). While the idea is theoretically compelling, empirical work has produced mixed results, with strong results in favor of (Chouinard & Clark, 2003; Hirsh-Pasek et al., 1984) and against (Brown & Hanlon, 1970; Hayes & Ahrens, 1988).

However, much of this prior work investigates tuning as an overt effort on behalf of parents or tuning with respect to content words, with less examining the potential role of low-level syntactic influence (Hoff, 2003). Yurovsky et al. (2016) presents just such an examination, demonstrating using Bayesian hierarchical modeling that parents align to their children according to their particular language usage at the level of syntactic categories. This paper extends their model by applying it to a new socioeconomically diverse sample of families (Goldin-Meadow et al., 2014) and leveraging the model's alignment estimates to predict language development outcomes.

The analysis presented here largely replicates the findings of Yurovsky et al. (2016), showing strong alignment effects for both parents and their children, a substantial age effect for baseline useage in children, and a trending negative effect of age on alignment for parents. Moreover, we demonstrate that these alignment estimates have substantial power in predicting language development outcomes, even in the presence of demographic features such as gender and socioeconomic status. As such, these results serve to further the linguistic tuning hypothesis, showing that alignment is a robust effect that appears to have a relationship with language development independent of demographic correlates.

References

Brown, R. W., & Hanlon, C. (1970). Derivational Complexity and Order of Acquisition in Child Speech. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 11–53). New York.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1).

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, *30*(3), 637–669.

Dunn, L. M., & Dunn, L. M. (n.d.). Peabody Picture Vocabulary Test–Third Edition.

Goldin-Meadow, S., Levine, S. C., Hedges, L. V., Huttenlocher, J., Raudenbush, S. W., & Small, S. L. (2014). New evidence about language and cognitive development based on a longitudinal study: Hypotheses for intervention. *American Psychologist*, *69*(6), 588–599.

Gomez, R. L., & Gerken, L. (1999). Artificial gram-

mar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135.

Hayes, D. P., & Ahrens, M. G. (1988). Vocabulary simplification for children: a special case of 'motherese'? *Journal of Child Language*, *15*(2), 395–410.

Hirsh-Pasek, K., Treiman, R., & Schneiderman, M. (1984). Brown & Hanlon revisited: mothers' sensitivity to ungrammatical forms. *Journal of Child Language*, *11*(01), 81–88.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*(5), 1368–1378.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, *20*(3), 61–27.

MacWhinney, B. (2000). The CHILDES Project. *Computational Linguistics*, *26*(4), 657–657.

Mayor, J., & Plunkett, K. (2010). A statistical estimate of infant and toddler vocabulary size from CDI analysis. *Developmental Science*, *14*(4), 769–785.

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8- and 12-year-old children and adults. *Journal of Experimental Psychology: General*, *144*(2), 447–468.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, *17*(02), 357–374.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). Linguistic inquiry and word count: LIWC2015. Austin, TX: Pennebaker Conglomerates.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Snow, C. E. (1972). Mothers' Speech to Children Learning Language. *Child Development*, *43*(2), 549–565.

Sokolov, J. L. (1993). A local contingency analysis of the fine-tuning hypothesis. *Developmental Psychology*, *29*(6), 1008–1023.

Spivey, M. J., & Dale, R. (2006). Continuous Dynamics in Real-Time Cognition. *Current Directions in Psychological Science*, *15*(5), 207–211.

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-Directed Speech Facilitates Word Segmentation. *Infancy*, *7*(1), 53–71.

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of School Outcomes Based on Early Language Production and Socioeconomic Factors. *Children and Poverty*, *65*(2), 606–621.

Weisleder, A., & Fernald, A. (2013). Talking to Children Matters. *Psychological Science*, *24*(11), 2143–2152.

Yuan, S., & Fisher, C. (2009). "Really? She Blicked the Baby?": Two-Year-Olds Learn Combinatorial Facts About Verbs by Listening. *Psychological Science*, *20*(5), 619–626.

Yurovsky, D., Doyle, G., & Frank, M. C. (2016). Linguistic input is tuned to children's developmental level. In *Proceedings of the annual meeting of the cognitive science society* (pp. 2093–2098).