The Development of Canonical Proportion as a Function of Community, Multilingualism, and Target Language's Syllable Complexity

Kai Jia TEY[1], Sarah WALKER[2], Amanda SEIDL[3], Camila SCAFF[1,4], Loann PEUREY[1], Bridgette L. KELLEHER[5], William N. HAVARD[6], Lisa R. HAMRICK[7], Pauline GROSJEAN[2], Margaret CYCHOSZ[8], Heidi COLLERAN[9], Marisa CASILLAS[10], Elika BERGELSON[11], Kasia HITCZENKO[12], Alejandrina CRISTIA[1]

[1] Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Etudes Cognitives, ENS, EHESS, CNRS, PSL University, Paris, France

[2] School of Economics, University of New South Wales, Sydney, Australia

[3] Department of Communication Sciences and Disorders, University of Delaware, Newark, DE, USA

[4] Institute of Evolutionary Medicine, University of Zurich, Zürich, Switzerland

[5] Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA

[6] Laboratoire Ligérien de Linguistique, University of Orléans, Orléans, France

[7] Department of Psychology, University of South Carolina, Columbia, SC, USA

[8] Department of Linguistics, University of California Los Angeles, Los Angeles, CA, USA

[9]BirthRites Lise Meitner Research Group, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[10] Department of Comparative Human Development, University of Chicago, IL, USA

[11] Department of Psychology, Harvard University, MA, USA

[12]Department of Computer Science, The George Washington University, Washington, DC, USA

Acknowledgments

This work was funded by the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award; European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ExELang, Grant agreement No. 101001095). We have no known conflict of interest to disclose. ChatGPT was used for minor language refinement. No AI tools were used for data analysis, study design, or interpretation.

Correspondence concerning this article should be addressed to Kai Jia TEY, Laboratoire de Sciences Cognitives et de Psycholinguistique (ENS, EHESS, CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, 29, rue d'Ulm, 75005 Paris, France. Email: kaijiatey@gmail.com.

Keywords: canonical proportion; language syllable complexity; multilingualism; community; vocal development; language development; canonical vocalizations; long-form recordings

Abstract

This study investigates the development of canonical proportion (CP), an indicator of vocal maturity, across diverse language and environmental contexts. Using the Speech Maturity Dataset (SMD) comprising 369 children, aged 0;2-6;4, across ten different languages and cultures, we explore the influence of multilingual exposure, language syllable complexity, and community type (industrialized, non-industrialized) on early phonological development. We find that monolingual children display higher CP measures than their multilingual peers. Additionally, CP is higher for children learning languages with simple syllable complexity than those with more complex syllables. We also find no significant differences in the CP trajectory of children from industrialized versus non-industrialized communities. We integrate these findings focusing with the broader literature on early language development, and highlight the importance of diversifying our participant samples to better understand the complex relationship between various dimensions of language exposure, social environment, and language development.

Introduction

Although language is universal among humans, a host of research suggests meaningful individual and group variation that can potentially shed light on the factors and processes involved in typical and atypical language development, including vocal development. Fine-grained phonetic measures suggest vocal development continues well beyond early childhood (Goffman, 1999; Nip & Green, 2013). Recent work suggests that canonical proportion (CP, introduced below) is one measure of vocal development that can be deployed in a large age range at scale, providing us with sufficient statistical power to study potential predictors, such as gender and linguistic diversity (Cychosz et al., 2021; Ott & Cychosz, under review). The present study builds on a recently released large dataset, the Speech Maturity Dataset (SMD, Hitczenko et al., submitted), which builds upon previous work (Cychosz et al., 2021; Hitczenko et al., 2023; Semenzin et al., 2021), and includes data from hundreds of children growing up exposed to various language and cultural backgrounds. Our primary goal is to harness variability in the Speech Maturity Dataset to understand how several aspects of language environments may relate to the development of CP.

Canonical Proportion in Early Vocal Development: Background and Previous Research

Our work builds conceptually on a strong research tradition examining the development of babbling during the first year of life. According to this research, canonical vocalizations, containing clear consonant-vowel transitions, are linked to the maturation of the articulatory system (Oller & Eilers, 1988). Canonical vocalizations begin to appear in infants' babble at around seven months of age, with canonical babbling ratios (the ratio of canonical syllables relative to other speech-like syllables in infants' meaningless babbling) reaching about 15% by 10 months of age (Oller et al., 1998, 1999). These early measures of vocal development have been linked to developmental outcomes. Studies show that infants

who go on to develop speech, language, or hearing impairment show lower canonical babbling ratios (e.g., hearing loss: Jung & Houston, 2019; reading disorder: Lambrecht Smith et al., 2010; Autism Spectrum Disorder: Patten et al., 2014; Paul et al., 2011; Yankowitz et al., 2022). Similarly, canonical babbling onsets (the age at which canonical syllables emerge) predict later vocabulary (Lieberman et al., 2024; Nathani et al., 2007; Oller et al., 1999).

While canonical babbling ratios and canonical babbling onsets have only been studied in infants' babbling, Cychosz et al. (2021) proposed an extension that enables researchers to study a conceptually related aspect of vocal development across infants' and young children's vocalizations (whether early babbling or later-developing meaningful speech). In this approach, vocalizations were arbitrarily divided into small audio clips classified into three types. The first type included all non-speech-like vocalizations (like crying and laughing), which will not be studied further here. The other two types of speech-like classes were canonical (containing clear consonant-vowel or vowel-consonant transitions) or non-canonical (speech-like but lacking a clear consonant-vowel or vowel-consonant transition). CP is defined as the ratio of audio clips that are classified as canonical among all those that are classified as speech-like (canonical or non-canonical).

To date, there have been only four studies using CP, which we summarize next. Two cross-linguistic investigations have examined parts of the dataset used in this study, with each successive work incorporating more data. Cychosz et al. (2021) analyzed 52 infants aged 1 to 36 months, who were growing up in a variety of communities, as monolinguals (English, Tsimane', Yélî Dnye, Tseltal) or bilinguals (Quechua/Spanish, English/Spanish, Yélî Dnye and other languages). The authors found that, similarly to babbling onset ratios, CP in these data reached about 15% by 10 months of age. However, by employing CP, they could ascertain that this percentage continued to increase in toddlerhood, reaching about 40% by 36 months. The authors argued that CP thus defined is a broader indicator of vocal development

that remains relevant beyond the babbling stage. In addition, they did not visually detect large differences across groups, which they interpreted as a sign that CP may be resilient to differences across languages and/or communities. Perhaps due to sample size limitations, they did not explicitly test for a difference between monolinguals and the others.

Building on Cychosz et al. (2021)'s dataset, Hitczenko et al. (2023) extended coverage by adding 77 children (total N = 129 children) from three additional communities. By extending the age range to 6 years, they documented increases in CP beyond toddlerhood reaching up to ~80%, lending further credence to Cychosz et al.'s argument that CP may continue to track vocal development beyond the pre-linguistic period. With increased power, Hitczenko et al. (2023) also revealed differences in CP based on the typological properties of the language(s) children were learning and whether children were growing up in an industrialized community. However, they did not investigate potential differences as a function of multilingualism. Since our analyses include their dataset, we defer detailed discussion of their results.

A third study focused on North American English-learning children. In addition to demonstrating high levels of convergence across laboratory coding and citizen scientist-based approaches, Semenzin et al. (2021) documented age-related CP increases among low risk controls aged 4 to 18 months, whereas children with Angelman syndrome (a neurogenetic disorder) aged 11 to 53 months exhibited age-related decreases. This aligns conceptually with the work summarized above linking lower canonical babbling ratios with atypical language development (e.g., Patten et al., 2014).

Additionally, recent research by Ott and Cychosz (under review) examined CP in 130 English-learning children at mean age three years, whose CP ranged between 20% and 76%. They observed a positive correlation between CP and age in their sample of 28 to 49 months old, with CP ranging from reaching an average 59% by the end of the age range. Their study

also demonstrated that CP measured at age three years significantly predicts multiple widely used standardized assessments of speech and language, which the children completed one year later at approximately four years old, including consonant articulation (Goldman Fristoe Test of Articulation-2), vocabulary size (Peabody Picture Vocabulary Test-4), phonological awareness (Comprehensive Test of Phonological Processing-2), and phonological working memory (non-word repetition tasks). Their findings suggest that CP may measure children's speech development beyond the babbling stage and could serve as a potential indicator of later speech and language development.

Potential factors that may influence the development of CP

The emergent literature focusing on CP suggests that, although there are systematic increases in CP with age, there may be meaningful individual (Ott & Cychosz, under review) and group variation (Cychosz et al., 2021; Hitczenko et al., 2023; Semenzin et al., 2021). Here, we study three factors that vary in our dataset (the Speech Maturity Dataset or SMD; Hitczenko et al., under review) and that reflect language and environmental diversity: multilingual exposure, syllable complexity, and community. In this section, we justify why these three factors might (not) influence the trajectory of CP.

Multilingual exposure

Should monolingual status affect CP? Since CP reflects the proportions of well-formed syllables in children's speech, it serves as a potential indicator of whether multilingual exposure influences the rate of phonological development. One long standing hypothesis in the field of language development has postulated that multilingual exposure might delay early speech milestones (a discussion in Fibla et al., 2021), because of at least two reasons: First, the presence of multiple linguistic systems may introduce cognitive and linguistic complexity and thus confuse the learners; and second, all else equal, multilinguals may receive less input in each of their languages than corresponding monolinguals, which

could influence how quickly stable speech forms emerge (in perception and/or production). Although there is a long controversy in the field of lexical acquisition (with some work suggesting delays, and others showing none; Bialystok et al., 2010; Byers-Heinlein, 2013; Hoff et al., 2012; Oller et al., 2007; Pearson et al. 1993), the bulk of the evidence suggests that there may not be any difference between monolinguals and multilinguals. However, there is less work pertaining to phonological structure. By examining CP in monolingual and multilingual children, we aim to determine whether multilingual exposure influences not just vocabulary size, but the very structure of early speech patterns. Evidence of this comes from both specific and broader overall multilingualism studies. Oller et al. (1997) examined 29 English-Spanish bilingual infants and found they exhibited similar canonical babbling ratios as 44 monolingual infants. In contrast, more recently, Bergelson et al. (2023) took a wider perspective, analyzing 1,001 children's speech-like vocalizations using data collected with LENA, a wearable device that records daylong, naturalistic language input and output. The sample included children (2 to 48 months) from diverse linguistic backgrounds, from both industrialized and non-industrialized communities, with some children at risk of atypical development. Despite the sizable power in their study, they found that the number of speech-like vocalizations children produced did not vary as a function of multilingual status (but see Zheng et al., 2022). Admittedly, no previous study looked at CP specifically, a gap we sought to fill.

Syllable complexity

Some previous studies have demonstrated that ambient language influences early phonological development, focusing for instance, on the phonetic variability in canonical vocalization, as well as the early phoneme and syllable sequences in infants (Andruski et al., 2014; Boysson-Bardies et al., 1989; Boysson-Bardies & Vihman, 1991; Levitt & Wang, 1991; Poulin-Dubois & Goodz, 2001; Sundara et al., 2020). Given that SMD contains data

from several different languages, but not enough in each, we sought to employ a typological classification that systematically differs across languages, and which may directly influence the various types of syllable structures that children are exposed to, since syllables are a foundation of speech production. For this, we turned to Maddieson (2013), who categorized languages as a function of the types of syllables allowed: a simple syllable complexity if only open syllables were allowed ((C)V); moderately complex syllable complexity if additionally some codas and complex onsets were allowed ((C)(C)V(C)); and complex syllable complexity if additional syllable shapes were permissible ((C)(C)(C)V(C)(C)(C)). The details of Maddieson's categorization will be discussed in the Methods section.

How may a language's syllable complexity status influence early phonological development? One possibility is that when only a small number of syllable shapes are allowed, as in simple syllable complexity languages, children have more exposure to the same template, which is also the easiest template for the young infant to produce. In contrast, languages with more complex syllable complexity introduce more phonotactic variability, requiring children to develop greater oral-motor control before achieving stable canonical transitions. Potentially, as infants begin to babble, their caregivers more easily recognize their babble as real words since there are fewer mismatches with the target word due to, for example, dropped codas or simplified clusters. Thus, regarding simple exposure accumulation and social feedback, one expects children learning simple syllable complexity languages to develop CP faster. Using a subset of the SMD, Hitczenko et al. (2023) found a significant effect of syllable complexity on CP. However, the strongest difference was between moderate and the other types, which challenges the assumption that syllable complexity linearly influences CP development. Similarly, Lee et al. (2018) investigated the canonical babbling ratio in long-form recordings from 21 infants learning English (a complex syllable complexity language) in the United States and Chinese (a moderate one) in both the

United States and Taiwan, hypothesizing that the lower level of syllable complexity found in Chinese may affect outcomes. Children learning Chinese showed numerically higher canonical ratios, although the difference was not statistically significant, which may indicate insufficient power. Together, these previous findings invite further attention to the input language's syllable complexity classification.

Communities

In this study, we follow previous work by adopting a simple first approach to classifying communities (Hitczenko et al., 2023; Cristia, 2023). communities are classified as non-industrialized if they have small-scaled populations, subsistence-based economies, and limited access to formal education and healthcare, in contrast to industrialized communities, which have a market-based economy and wide access to formal education and healthcare services. SMD contains data from children growing up in 10 communities. Based on this classification, most communities in SMD were considered non-industrialized. Several studies in children's lexical development have reported that children in non-industrialized communities have smaller vocabularies and/or lower language scores than their industrialized counterparts (e.g., Ma et al., 2021; Vogt et al., 2015; see also Ma et al., 2023). However, other studies have found comparable language development trajectories between children from non-industrialized and industrialized communities (Casillas et al., 2020, 2021). Moreover, CP may follow a different developmental trajectory. Unlike vocabulary growth, which depends mainly on input exposure and cognitive processing (Snedeker et al., 2012), vocal production may be less sensitive to aspects of the input varying across cultures (e.g., child-directed speech quantity; Cristia, 2023; Ma et al., 2021). That said, there may be other environmental factors that vary across communities (such as social interaction styles, and background noise), which may in their stead play a role in CP development.

This raises the question: how might community differences influence CP? The two previous studies on which we build reached opposite conclusions. Based on the first 52 children, Cychosz et al. (2021) commented on the lack of salient visual differences in CP across the various communities (each represented by only 3-16 children). In contrast, by increasing the sample size to 129 children and attempting a comparison based on a dichotomic non-industrialized/industrialized distinction, Hitczenko et al. (2023) found significant variability: Children in non-industrialized communities showed higher CPs than those in industrialized communities but similar developmental trajectories, as there were no interaction effects with age. Here, we revisit the question in a cumulative fashion, since SMD includes more than double the number of participants.

The present study

The emergent research on CP suggests meaningful age-based variation, but further work is needed to understand how children's language and environmental experiences relate to that measure. Our study harnesses variability in SMD to measure how CP varies by three experiential factors: multilingualism, syllable complexity, and community characteristics. Since our dataset includes data analyzed in Hitczenko et al. (2023), we hypothesized, based on their findings, that CP might vary by syllable complexity and community type (non-industrialized vs. industrialized). Our expanded dataset allows us to investigate further whether these differences hold and explore potential explanations for CP variability across community context. In conceptual terms, we believe the present research can contribute relevant information to deepen our understanding of the factors potentially influencing vocal, phonological, and language development more broadly.

Methods

Reproducibility of analyses has been ensured using a repository (HIDDEN FOR REVIEW).

Dataset

Our data come from the Speech Maturity Dataset (SMD; Hitczenko et al., under review). Children were individually recorded using an unobtrusive wearable recording device, resulting in long audios (e.g., 15 consecutive hours), which were analyzed either manually or using state-of-the-art software (called VTC, Lavechin et al., 2020) to automatically identify sections of the audio in which the child or others vocalized (i.e., babbled, cooed, or spoke). These vocalizations were sampled using different methods, including: key child vocalization sampling (extract audio sections attributed to the key child), loudness sampling (extract audio sections based on the amplitude profile, so as to be independent from VTC), and female adult vocalization sampling (extract audio sections attributed to female adults). Once sampled, these audio sections were cut into short ~ 500ms clips to preserve participants' sensitive information and voice identity. The clips were then uploaded to a citizen science platform. For the vast majority of the data, the platform was Zooniverse, and the project was called the Maturity of Baby Sounds (https://www.zooniverse.org/projects/laac-lscp/maturity-of-baby-sounds). In this case, several thousand non-expert individuals contributed to crowdsourcing classifications after minimal training. The data from the 52 participants from Cychosz et al. (2021) came from a different citizen science platform, which is no longer available. In all cases, at least three citizen scientists classified each individual clip into different categories: canonical, non-canonical, laughter, crying or none of the above. Considering only clips in which majority annotators agreed on the label, Hitczenko et al. calculated CP for each child by dividing the total number of clips that had been classified as canonical by the sum of the number of clips classified as

either canonical or non-canonical. They did this for every combination of recording date and sampling method. For instance, a child who was recorded on two separate days, and had key-child as well as female-adult vocalizations separately sampled from each day, would have four CP values associated with their participant ID, only two of which - the two key-child CP values - are relevant to the present study. For a more comprehensive description of the data collection and processing involved in the SMD, refer to Hitczenko et al. (under review).

From the 651 CP measures available in SMD, we excluded ten CP measures from children with atypical language development (Semenzin et al., 2021), one CP measure lacking gender information (due to the sampling method), three CP measures with no age information, two measures lacking key-child speech-like vocalizations (which made CP calculation impossible), and 69 CP measures resulting from a different sampling method that was not the focus of our study (the female-adult sampling approach). In the end, we analyzed 566 CP measures.

Participants

We analyzed data from 371 unique children (193 boys; some children contributed data from > 1 day) aged 2 to 76 months. These children came from a variety of linguistic backgrounds, representing ten different corpora. Table 1 provides an overview of the characteristics of each corpus.

Table 1

Summary of Corpus Characteristics

Corpus	Language_Spoken	Number of Children	Age_Range	Multilingual Status	Syllable_Complexity	Community Type
Papua New Guinea	Yélî Dnye	46	2-76	mostly monolingual	simple	non-industrialized
Bolivia	Tsimane'	41	5-70	monolingual	moderate	non-industrialized
France	French	10	10-11	mostly monolingual	complex	industrialized
Mexico	Tseltal	10	1-36	monolingual	complex	non-industrialized
USA-Indiana	English	10	4-53	monolingual	complex	industrialized
USA-New York	English	10	7-17	monolingual	complex	industrialized
Solomon Islands	Solomon*	198	4-48	multilingual	NA	non-industrialized
Vanuatu	Vanuatu*	40	5-51	multilingual	NA	non-industrialized
USA-California	English & Spanish	3	3	multilingual	NA	industrialized
Bolivia	Quechua & Spanish	3	24	multilingual	NA	non-industrialized

Note. **Solomon* consists of languages including Roviana, Avaso, Babatana, Marco, Marovo, Pidjin, Sengga, Simbo, Sisinga, Ughele, Vaghua, and Varisi

**Vanuatu* consists of languages including Bislama, Venen Taut, Petarmul, Neverver, Uripiv, Vinmavis, Novol, Epi, Nah'ai, Paama, Ninde, Tautu, French, Pinalum, Malo, Rano, Tauta, Santo Language, Ambae, Maevo, South, Atchin, and Tempun

SMD participants are classified as monolingual or multilingual following Bergelson et al. (2023): monolinguals are reportedly exposed to only one language (29.6% in the dataset), and otherwise they are multilinguals (i.e., non-monolinguals) regardless of the number of languages.

Monolingual children were further categorized based on the syllable complexity of their input language. Although it would have been possible to classify multilingual children as the highest syllable complexity of input languages, we thought this may introduce additional noise due to inconsistent exposure levels across languages. Classification followed Maddieson's (2013) simple, moderate, and complex syllable complexity languages as follows. In languages classified as having a SSC, syllable structure is restricted to (C)V (C: consonant; V: vowel) sequences, allowing no onset consonant clusters, and only V and CV as permissible syllables. This category is relatively rare, representing only 12% of the languages studied by Maddieson. In our data, Yélî Dnye is the only language representing this category.

In contrast, languages classified as having moderate syllable complexity allow single codas and/or consonant clusters (typically involving an approximant or liquid) in onsets. This includes syllable types such as VC, CVC, CCV, and CCVC, in addition to the simpler

structures V and CV. Around 57% of the languages studied by Maddieson are in this category. Tsimane' is the only language classified as moderate syllable complexity in SMD.

Finally, languages with complex syllable complexity allow more intricate consonant clusters in both onsets and codas. Some examples of syllable structures that are allowed in languages with complex syllable complexity are V, CV, VC, CVC, CCV, CCVC, CVCC, CCVCC, CCCVCC, CCCVCCC. The remaining 31% of languages were classified by Maddieson in this category. In SMD, French, English, and Tseltal are classified as having complex syllable complexity.

As for the last factor focused in this study, communities were classified as industrialized or non-industrialized (Hitczenko et al., 2023; Cristia, 2023). Our dataset includes four corpora from industrialized communities and six corpora from non-industrialized communities (see Table 1 for details). Although we acknowledge that the actual classification of communities in the real world is a complex and multifaceted process, the simple classification serves as a starting point for our analysis.

Analyses

All analyses were conducted in R version 4.3.2 (R Core Team, 2024), using the stats (R Core Team, 2024) and car (Fox & Weisberg, 2019) packages for statistical analysis, and the ggplot2 (Wickham, 2016) package for data visualization. To better match age distributions and in the presence of multicollinearity, we examined the predictive value of multilingual exposure, native language syllable complexity, and community type in three separate mixed-effects logistic regression models, with CP as the dependent variable. Weighted analyses were employed to account for the differences in number of clips available for each child and ensure accurate representation in the models. Specifically, each CP measure was weighted by the total number of clips it was derived from, i.e., excluding those labeled as junk, NA, no majority label, and non-speech. The linear and quadratic terms of age

(both z-scored) were also included, in interaction with the main effect being studied. As our study builds on Hitczenko et al. (2023), with a larger dataset, we followed many of their statistical analysis approaches, including the inclusion of the quadratic term for age. However, we didn't include it blindly, as we also observed clear evidence of a quadratic age effect in visual inspection, which reinforced our decision to incorporate it into our models. We fitted the models using a top-down approach, starting with the most complex model, which included all predictors and interactions, and then comparing it to reduced models by removing non-significant terms step by step based on AIC and likelihood ratio tests. Multilingual exposure, syllable complexity, and community were treated as categorical variables, with monolinguals, simple syllable complexity, and non-industrialized communities set as reference levels, respectively. Re-levelling was performed in relevant analyses to facilitate specific pairwise comparisons among different categories, as described in the Results section. Child_id was treated as a random variable nested within the corpus to account for individual and corpus-level variability. Type 3 ANOVA tests were performed to confirm the significance of main effects and interactions identified in the regression models. Further model diagnostics were performed to validate the tests' assumptions.

Results

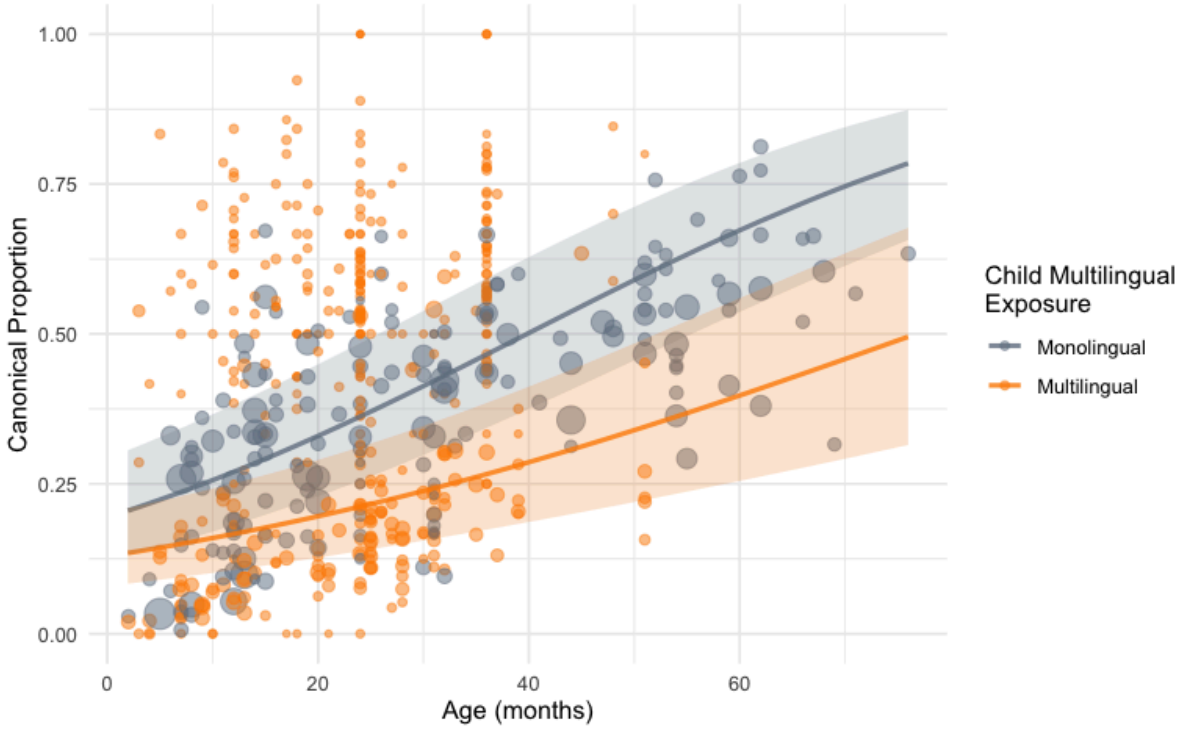
Multilingual exposure

To investigate whether CP varies as a function of multilingual exposure, we analyzed data from all participants, namely 371 children, spanning the entire age range (2-76 months). We fit a mixed-effects logistic regression model to test the effects of multilingual status, age, and their interaction on CP: *CP ~ age*multilingualism + age^2*multilingualism + (1|corpus/child_id).*

The results revealed a significant main effect of multilingualism (Estimate = -.74, SE = .17, z = -4.34, $p < .001$), indicating that monolingual children demonstrated higher CP measures compared to multilingual ones. At the mean age of the sample (25.6 months), CP is estimated at 0.38 for monolinguals and 0.22 for multilinguals, leading to a difference of 72.7% higher CP for the former compared to the latter. Both linear and quadratic terms for age were significant predictors of CP (Estimate = .48, SE = .05, z = 8.88, $p < .001$; Estimate = -.11, SE = .04, z = -2.74, $p < .01$). These results suggest that CP increases with age overall (as indicated by the positive linear term), but the significant negative quadratic term suggests that this increase slows down with age. Notably, no significant interaction between age and multilingualism was observed, suggesting that age-related trends were similar enough across monolingual and multilingual groups. Figure 1 shows the CP distribution in monolingual and multilingual children across the children's age. To further assess the robustness of these findings, we conducted an additional analysis where we controlled for and balanced the number of children from Solomon Islands, see Supplementary Materials Section 1. This was done to ensure that differences in CP were not driven by population imbalances. The results remained consistent. Additionally, we performed an analysis subsetting to shared age ranges, see Supplementary Materials Section 2, which confirms the significant effect of multilingualism but finds no evidence for a quadratic effect of age.

Figure 1

Canonical proportions by Age and Multilingual Exposure (Full sample)



Note. The regression line represents the fitted model, and the shaded bands surrounding the line represent the 95% confidence intervals. Each data point represents a single child, with point size representing the total number of vocalizations contributed by that child (larger points represent children who produced more vocalizations).

Additional diagnostic tests were conducted to assess the model fit. Random effect diagnostics showed moderate variance ($\sigma^2 = 0.36$ and $\sigma^2 = 0.48$), supporting the inclusion of the random effects. A likelihood ratio test comparing the full model to a reduced model (excluding multilingualism) showed a significantly better fit for the full model ($X^2 = 20.01$, df = 3, $p < .001$). Similarly, a likelihood ratio test comparing the full model to a reduced model excluding the quadratic age term also showed a significant improvement in model fit ($X^2 = 9.08$, df = 2, $p < .05$), confirming the inclusion of the quadratic age term.

Overall, these results suggest that CP increased with age, with growth slowing over time; and that monolingual children consistently exhibited higher CP than multilingual children.

Syllable Complexity

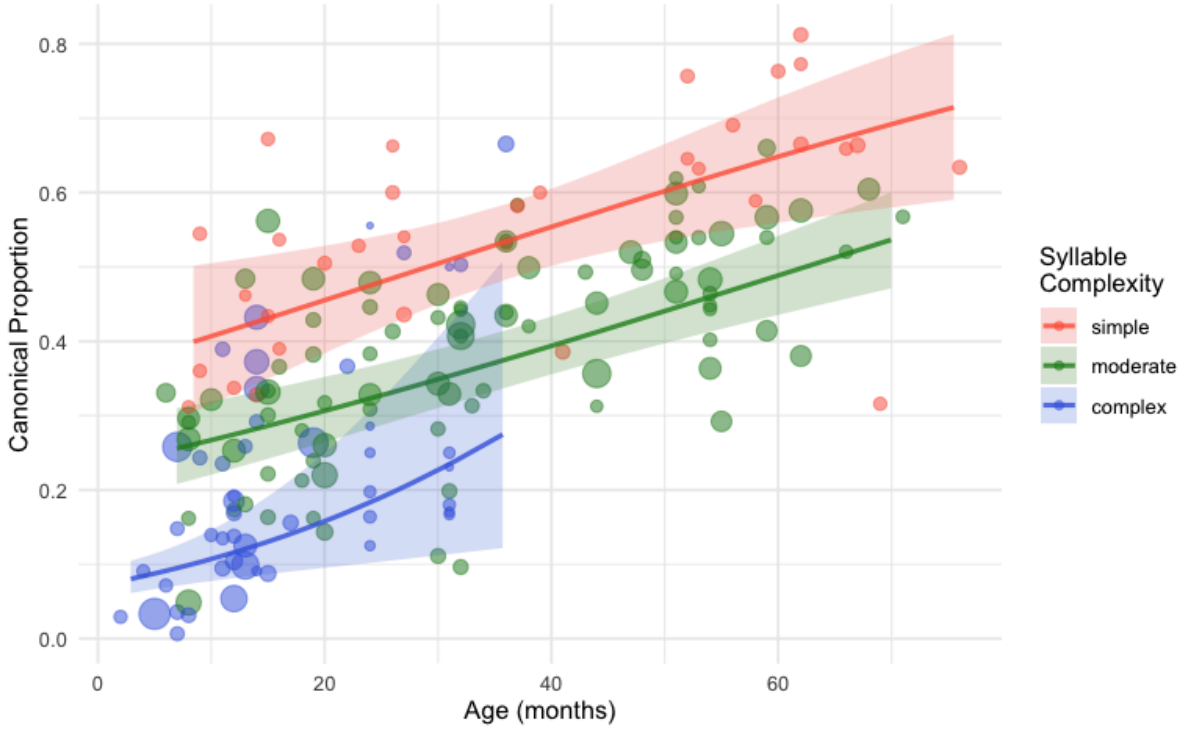
Second, we investigated the relationship between syllable complexity and CP among monolingual children. Since this analysis focuses only on monolingual children (as we do not have detailed information on languages spoken by each child in multilingual settings), we included 110 monolingual children (31 learning a simple syllable complexity language, 41 moderate, 38 complex). This selection excluded 261 children from the analysis. We fit a mixed-effects logistic regression model to test the effects of syllable complexity, age, and their interaction on CP: $CP \sim age*syllable\_complexity + age^2*syllable\_complexity + (1|corpus/child\_id)$.

This analysis revealed a significant main effect of syllable complexity on CP. Specifically, children learning a language of moderate syllable complexity (Estimate = -.70, SE = .15, z = -4.81, $p < .001$) and complex syllable complexity (Estimate = -.93, SE = .27, z = -3.43, $p < .001$) exhibited significantly lower CP compared to the reference category, namely children learning a language with simple syllable complexity, who exhibited the highest CP. Re-leveling syllable complexity with "moderate" as the reference category showed that the difference between moderate and complex syllable complexity was not statistically significant (Estimate = -.23, SE = .26, z = -0.88, $p = .379$). At the mean age of the sample (29.8 months), CP is estimated at 0.5 for simple, 0.35 for moderate, and 0.23 for complex, leading to a difference of 42.9% higher CP for the simple versus moderate, and 52.2% for moderate versus complex. As in the multilingual model discussed in the previous section, the linear age predictor had a significant positive effect on CP (Estimate = .36, SE = .12, z = 2.97, $p < .005$), and the interactions were not significant (all $p > .05$). However, unlike

in the previous analysis, the quadratic term for age did not explain significant variance. This suggests that disaggregating monolinguals by syllable complexity reveals less of a plateau in older age groups than in the combined analysis represented in Figure 1. Figure 2 shows monolingual children's CP as a function of age and their native language's syllable complexity level.

Figure 2

Canonical proportions by Age and Syllable Complexity in Monolingual Children



Note. The regression line represents the fitted model, and the shaded bands surrounding the line represent 95% confidence intervals. Each data point represents a single child, with point size indicating the total number of vocalizations contributed by that child (larger points represent children who produced more vocalizations).

To evaluate model fit, we conducted additional diagnostic tests. Random effect diagnostics showed low variance ($\sigma^2 = 0.27$ and $\sigma^2 < 0.001$). The likelihood ratio test comparing the full model to a reduced model (excluding syllable complexity) showed a significantly better fit for the full model ($X^2 = 43.97$, df = 6, $p < .001$).

One concern is that the three groups of monolinguals overlap only in a subset of the age range. In an analysis reported on in Supplementary Materials Section 3, we subset to the shared age range (8-36 months). Although the sample size was not much smaller (total N = 70 children; 16 simple syllable complexity, 24 moderate, 30 complex), the main effect of syllable complexity failed to achieve statistical significance and there was an age by complexity interaction. Visual inspection and a consideration of estimated sizes suggest to us that large enough sample sizes and age ranges are required to substantiate developmental differences as a function of syllable complexity with statistically significant tests.

Communities

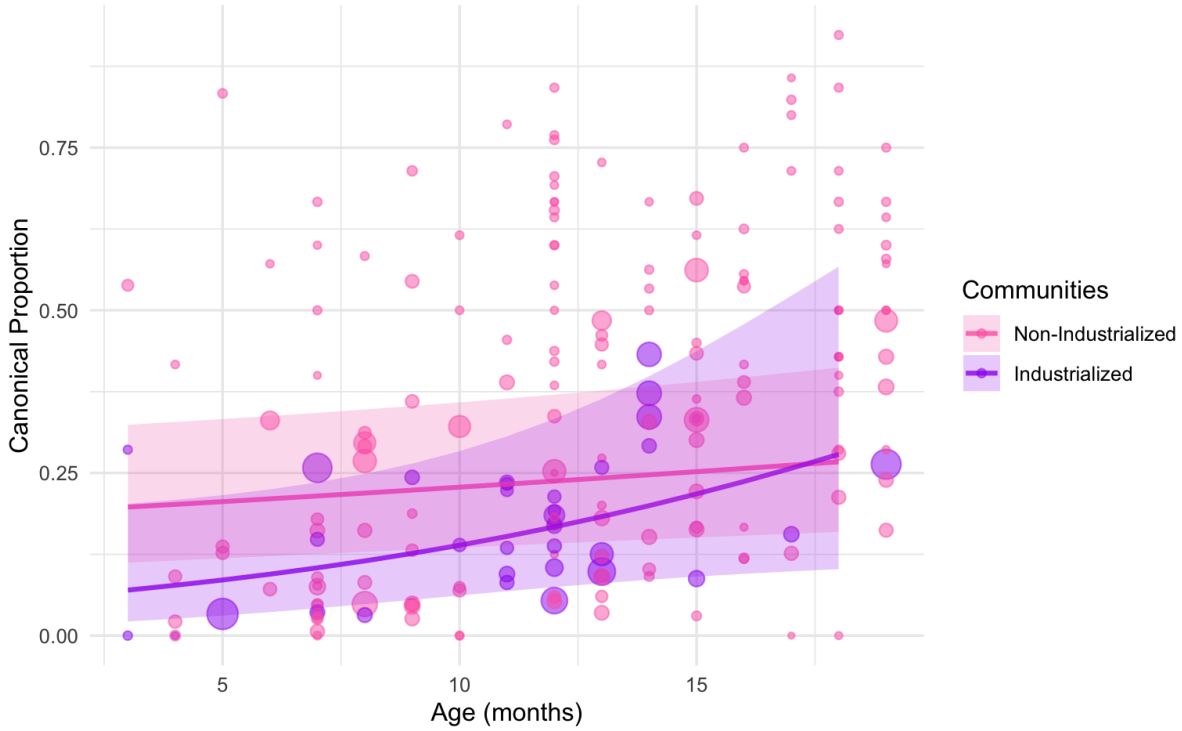
Last, we examined the relationship between community types (industrialized versus non-industrialized) and CP. Noticing an unequal age ranges between industrialized and non-industrialized groups, we subset data to children aged 3-19 months, resulting in 149 children (116 non-industrialized, 33 industrialized) being included in a mixed-effects logistic regression model: $CP \sim age*community + age^2*community + (1|corpus/child\_id)$.

The analysis revealed no significant main effects for the community (Estimate = -.42, SE = .57, z = -0.73, p = .46), suggesting no significant differences in CP measures between community types. Both the linear (Estimate = .11, SE = .05, z = 2.14, $p < .05$) and quadratic (Estimate = .14, SE = .06, z = 2.35, $p < .05$) terms for age were positively associated with CP, suggesting a developmental increase in CP over time that speeds up in this age range (3 to 19 months), unlike the significant plateauing observed in the analysis centered on multilingualism which spanned our full age range (3 to 77 months). No significant interaction was observed between community and age. Visual inspection of developmental trajectories illustrated in Figure 3 reveals that children from non-industrialized communities exhibited a numerically higher CP than industrialized children at the youngest ages, with this difference narrowing with age. Visually, this pattern goes against what we might expect based on

experiential factors. Given that industrialized children are typically exposed to more structured linguistic input, they would be expected to have higher CP. However, our results show the opposite, non-industrialized children start with higher CP, though the gap narrows with age.

Figure 3

Canonical proportions by Age and Community



Note. The regression line represents the fitted model, and the shaded bands surrounding the line represent 95% confidence intervals. Each data point represents a single child (N = 116 non-industrialized, N = 33 industrialized), with point size indicating the total number of vocalizations contributed by that child (larger points represent children who produced more vocalizations).

Additional diagnostic tests were conducted to assess the model fit and variability between groups. Random effect diagnostics showed moderate variance ($\sigma^2 = 0.68$ and $\sigma^2 = 0.46$), supporting the inclusion of the random effects. A likelihood ratio test comparing the full model to a reduced model (excluding community) failed to show a significantly better fit for the full model ($X^2 = 7.41$, df = 3, $p = .06$).

Discussion

The present study examined the relationship between CP and three factors: multilingual exposure, ambient language syllable complexity, and community. In SMD, monolingual children were shown to have higher CP measures than multilingual children. At the mean age of the sample (25.6 months), CP was estimated at 0.38 for monolinguals and 0.22 for multilinguals, meaning that monolinguals exhibited a 72.7% higher CP than multilinguals. We also found that children learning languages with simple syllable complexity exhibited the highest CP, followed by those learning languages with moderate and complex syllable complexity. At the mean age of the sample (29.8 months), CP was estimated at 0.50 for simple, 0.35 for moderate, and 0.23 for complex syllable complexity. This reflects a 42.9% higher CP for children learning simple versus moderate syllable complexity, and a 52.2% difference between moderate and complex syllable complexity. Interestingly, the difference in CP between children learning moderate versus complex syllable complexity was less pronounced. Finally, we failed to find a significant difference between children from non-industrialized and industrialized communities. Next, we discuss each of these effects.

Multilingual exposure

We found significant differences in CP development between monolingual and multilingual children, with a higher mean CP observed among monolingual children. While one possibility is that multilingual exposure affects CP through reduced input per language (Fibla et al., 2021), our findings also suggest that phonotactic complexity may play a role. Since we do not have detailed exposure data for multilingual children, it remains unclear whether CP differences stem from input quantity, linguistic structure, or a combination of both. To our knowledge, the only other empirical result partially aligned with this idea comes from one analysis reported by Zheng et al. (2022): In an Australian sample, monolingual

children produced more vocalizations than multilingual children at preschool, a difference that was not statistically significant after controlling for variables such as family income and home environment. Of note, previous studies reported no significant differences in babbling and speech-like vocalization patterns between monolingual and multilingual children (Oller et al., 1997; and Bergelson et al., 2023 respectively).

Another possibility is that there truly is a difference between monolinguals and multilinguals, which only our study could detect. To begin with, we could analyze a fairly large sample (compare our 171 multilingual to Oller's 29), which gives us more statistical power than others. That said, both Zheng et al. (2022) and Bergelson et al. (2023) included a fairly large sample, suggesting that power is not the only reason why we may have found a positive effect where they did not. Another difference is that those studies looked at the number of speech-like vocalizations, which may be conceptually distinct from CP. Metrics like CP, as a more fine-grained measure, may capture developmental differences that broader measures, like overall vocalization counts, do not detect. Due to the unequal distribution of vocalization counts in our dataset, we did not conduct further statistical analysis on this measure. Future research could test this directly by comparing monolinguals and multilinguals on both CP and vocalization counts. We would predict a difference between them in the former but not the latter metric.

Another possibility is that our apparent difference between monolinguals and multilinguals actually reflects a confounding factor. In our study, multilingual children were more likely to be learning languages with simpler syllable complexity. Does syllable complexity account for the observed differences? The answer is clearly no: Multilingual participants in our study spoke languages with simple syllable structures, which would typically predict higher CP. This expectation contradicts our observed pattern, suggesting that syllable complexity alone does not explain our findings.

Finally, it is worth considering that our results only begin to address a question that has received limited attention. It would be valuable for future research to explore the number and quantity of languages to which multilingual children in each sample are exposed. In our study, children from regions like Vanuatu and Solomon Islands are exposed to greater linguistic diversity, often encountering multiple languages (1-8 languages). The balance of exposure across these languages, whether evenly distributed or highly uneven, may influence developmental outcomes differently. Future studies with larger and more diverse samples are needed to better understand how input quantity, exposure equity, and linguistic factors like syllable complexity influence CP development in multilingual children.

Syllable complexity

Our findings suggest that a language's syllable complexity may affect CP development in monolingual children, provided a sufficiently large sample and age range is represented in the data. In analyses employing all monolingual data points, we observed that children learning languages with simple syllable complexity showed higher CP than those learning languages with moderately complex or complex syllables, which is in line with the hypothesis that languages with simpler syllable complexity allow earlier phonological development. Our results differ from those reported by Hitczenko et al. (2023), which were based on a subset of data, and found a different pattern, where moderate syllable complexity showed the strongest difference, while simple and complex syllable complexity patterned together. The discrepancy may be explained by the methodological difference in how multilingual children were treated. Hitckenzo et al. (2023) included both monolingual and multilingual children in their analysis of syllable complexity. They categorized multilingual children based on the idea that Austronesian languages tend to have simple syllable complexity. In contrast, we focused only on monolingual children when examining syllable complexity, to have a more precise description of each child's phonological exposure.

Another possible explanation is statistical power. Our dataset includes more children, resulting in higher statistical power, and allowing us to detect subtle effects of syllable complexity that may not be apparent in the smaller sample. However, even with the increased statistical power, we were still unable to detect a significant difference between moderate and complex syllable complexity, despite it being apparent visually. We believe further work is needed to better understand how syllable complexity affects vocal production and later development more generally. For instance, previous cross-linguistic research has shown that children learning languages with simpler syllable structures, such as Turkish (moderate syllable complexity), acquire phonological awareness faster than those learning languages with more complex syllable structures, such as English (Stringer, 2021). This aligns with our findings, suggesting that syllable simplicity may create a more favorable environment for early phonological development.

Interestingly, when we restricted the analysis to the 70 children aged 8-36 months old, we no longer found a significant main effect of syllable complexity on CP (see Supplementary Materials Section 3). This suggests that the previously observed syllable complexity effect may be sensitive to age range.

Communities

Unlike researchers studying a subset of these data (Hitczenko et al., 2023), we did not find significant differences between children growing up in industrialized versus non-industrialized communities. Hitczenko et al. (2023) reported higher CP in children growing up in non-industrialized communities. Here, thanks to a more comprehensive dataset, we were able to update that report, finding no significant overall differences as a function of community. A null result (as well as the previously reported finding of higher CP for children growing up non-industrialized than industrialized communities) is unexpected if CP is sensitive to speech input quantity, a possibility we raised as potentially accounting for

the difference we found between multilinguals and monolinguals. As we discussed above, while child-directed speech is known to support lexical development, its role in phonological development is less clear. If child directed speech influenced CP, and if children in non-industrialized communities received less child-directed speech (e.g., Cristia, 2023; Ma et al., 2021), they would be expected to show slower CP development. However, our results did not show such an effect, suggesting that CP may be less sensitive to variation in child-directed speech than other aspects of language development.

The absence of difference between children classified using the industrialized and non-industrialized binary is all the more surprising because there probably are many more ways in which children's development differs across the two population groups than prevalence of child-directed speech. For instance, infants in non-industrialized versus industrialized communities may be exposed to different background noise profiles, which may temporarily challenge infants, potentially due to underdeveloped attentional skills or the masking effects of noise on speech comprehension (Erickson & Newman, 2017). Despite these differences, the similarity in CP across these groups suggests that developmental trajectories remain resilient. We hope future theoretical work disaggregates the many factors that differ across so-called non-industrialized and industrialized communities (e.g., prevalence of child-directed speech, noise profiles, and others not discussed here), and helps us better account for group and individual differences in phonological development.

Age

The data from all three analyses can shed light on how CP changes with age. The fact that increases continue beyond 18 months aligns with previous conclusions that CP measure is not only interesting in the context of precursors to speech (Hitczenko et al., 2023). During the early stages of development, CP increases rapidly as children make significant progress in oral motor and phonological skills, thereby improving their ability to articulate complex

canonical forms (Anthony & Francis, 2005). In fact, our third analysis (Figure 3) suggests that in the 3- to 19-month range, CP is increasing rapidly, as evidenced by a positive and significant quadratic term. The first stage in which CP continues to increase goes on until at least 36 months of age (see Figure 2, regression line for complex syllable languages). However, as children reach more stable stages of speech production, the rate of CP growth slows. This slowing was reflected in the negative and significant estimate of the quadratic term of age observed in the first analysis (Figure 1). Although we did not carry out an analysis to pinpoint the age at which this occurs, visual inspection suggests that the plateau may be as early as 40 months.

Is it the case that these developmental trends vary as a function of the language's syllable complexity? Our analyses did not reveal a significant interaction, but we were constrained by not only power limitations but also age coverage. Notice that the oldest children learning languages with complex syllable complexity were only 36 months of age; for moderate and simple syllable complexity this was over 70 months. This difference in age coverage limits our ability to track longer-term trends in CP for children learning complex syllable languages. We hope that future studies will be better able to pinpoint potential differences in developmental patterns beyond 36 months. We hypothesize that developmental trends will be more pronounced for children learning languages with more complex syllable complexity. The acquisition of complex structures may require sustained effort and practice in their phonological acquisition, delaying their progress in canonical vocalizations. Thus, we predict the strongest differences in CP between complex and the other groups in the youngest ages. One issue that future work should look into more closely is the fact that children learning languages with simple or moderate syllable complexity tend to have a higher CP starting point than children learning complex syllable complexity languages, which is unexpected if differences truly are due to exposure. Moreover, starting from a higher point

reduces our ability to observe a noticeable slowdown in growth. Undoubtedly, much still remains to be learned about CP development in children learning a variety of languages. In the meantime, our dataset is already much larger than those used in previous studies, providing valuable insights into CP development across different linguistic environments.

A better understanding of CP differences

One crucial question for future research is understanding why CP differs among individuals and groups. What factors drive these differences? How do they emerge? Understanding these influences will help understand the role of CP in early language development. This and previous papers have built on research on phonological development to conceptualize CP as a measure of vocal development. But what precise conceptual aspect of vocal development does CP capture? One possibility is that CP captures differences in syllable duration, such that older children and children learning simple syllable complexity languages produce syllables that are shorter in duration as they develop greater speech motor control. Early in development, children's syllables tend to be longer and more variable due to immature motor coordination. As they gain more experience, their articulatory movements become more precise, reducing the transition time needed between consonants and vowels (Mahr et al., 2021; Nip & Green, 2013). This increased efficiency may contribute to shorter and more stable syllables (Goffman, 1999), hypothetically leading to higher CP measures.

Another possibility that we have not explored is that there are cross-linguistic and individual differences in the phonetic inventory. For instance, perhaps some of the variance in CP we observe relates to consonantal targets. Presumably, it will be easier for a child to achieve an adult-like CV transition when the C is a /d/ or /b/ than when the child tries to produce /ð/ or /ß/, which require finer oral motor control to generate turbulent flow at the place of constriction. Indeed, we grouped languages based on syllable structure, but there may be other phonetic, phonological, and structural differences which may affect CP. That

said, we do not think this possibility is likely. In broad terms, older children attempt more varied consonantal targets and have greater articulatory mastery. If increased articulatory complexity led to lower CP, then we would expect children over 4 years of age, in our data, to show lower CP than children aged about 1 year, at which age they typically only produce stops. However, this pattern is not observed, making this explanation unlikely.

Another possibility that does not fit the data relates to CP as a function of consonantal stretches: The idea being that in languages with greater syllable complexity, where onset clusters are allowed, the sections of the 500ms clip that reflect vowels or consonants are relatively shorter than in simpler syllables, and that this leads to a higher proportion of clips labeled as non-canonical. We think this is not a good explanation for several reasons. First, this would predict that children at ages where they start producing clusters would show lower CP, but this is not what we see in the data, where group differences between e.g. simple syllable complexity and complex syllable complexity languages are obvious even before 18 months, an age at which children do not readily produce many consonant clusters. Second, the 500ms duration almost certainly covers a full syllable, even for syllables with complex onsets and codas. Finally, our request to citizen scientists was for speech-like VC or CV transitions, which should not penalize children who produce more clusters.

Regardless of the possible interpretation, we also hope that future work develops language-specific benchmarks, for instance, by estimating CP for adolescents and adults, a direction that research using SMD has begun to explore (Hitczenko et al., under review).

Limitations

One important limitation of this study is the distribution of the syllable complexity and the community types within the SMD dataset. There is only one language representing each of the simple and moderate syllable complexity categories, and only three for the complex one. Additionally, we could only test the difference across industrialized and

non-industrialized communities in a narrow age range (3-19 months). Despite these limitations, this study analyzes the most extensive dataset available to date on early vocal development across diverse linguistic and community settings. These limitations highlight the need for larger and more balanced datasets in future research to enable a comprehensive analysis of how syllable complexity and community influence early language development.

Another limitation relates to the method used to categorize syllable complexity in this study, specifically the Maddieson (2013) framework. While this categorization is widely used, it may not adequately account for the unique syllable patterns found in all languages. For example, both Quechua and Spanish are classified as having moderate syllable complexity, yet Spanish permits complex consonant clusters (e.g., tres /tres/), while Quechua primarily follows a simpler (C)V(C) structure. Future research could consider analyzing the specific syllable patterns of each language to develop a more nuanced understanding of the relationship between syllable structure and CP development.

An additional limitation of SMD, and by extension our study, is the lack of detailed information on the specific language spoken by multilingual children and their extent of exposure to each language. This limitation is common in multilingual datasets, especially those involving large-scale data and under-studied languages. Additionally, our dataset reflects interdisciplinary collaboration and data reuse, which influenced our grouping strategy. For example, the data from Solomon Islands was part of a larger study combining economic games and a randomized controlled trial, limiting the collection of specific language details due to survey constraints. Furthermore, there are no standardized methods for quickly gathering multilingual exposure information that do not require lengthy (30-minute) surveys. This challenge makes it difficult to categorize children by factors thought to be key in bilingualism and multilingualism research, such as language dominance and proportion of exposure (Byers-Heinlein, 2015).

Conclusions

This study offers valuable insights into the potential impact of multilingual exposure, syllable complexity, and community factors on the development of CP in children's early language production. By analyzing a large and diverse dataset, which includes understudied languages and communities, our findings emphasize how universal developmental processes and language-specific characteristics impact early phonological development. Our findings suggest that multilingual exposure may explain some variability in CP development. Moreover, provided that a large enough sample and age range is represented, children exposed to languages with simple syllable complexity tend to have higher CP than those learning languages with moderate and complex syllable complexity, underscoring the influence of syllable structure of the ambient language on the development of speech production. Finding differences in phonological development as a function of these further emphasizes the importance of considering a range of linguistic and cultural contexts when studying language acquisition. Although an analysis based on a binary classification of communities into non-industrialized versus industrialized did not reveal significant differences, further research with more nuanced classification is needed to shed light on potential community-based differences. Ultimately, this study contributes to a nuanced understanding of the diverse pathways children take across languages during early phonological development.

References

Andruski, J. E., Casielles, E., & Nathan, G. (2014). Is bilingual babbling language-specific? Some evidence from a case study of Spanish–English dual acquisition. *Bilingualism: Language and Cognition*, *17*(3), 660–672. https://doi.org/10.1017/S1366728913000655

Anthony, J. L., & Francis, D. J. (2005). Development of phonological awareness. *Current Directions in Psychological Science*, *14*(5), 255–259. https://doi.org/10.1111/j.0963-7214.2005.00376.x

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramírez-Esparza, N., R. Hamrick, L., Marklund, E., Kalashnikova, M., Guez, A., Casillas, M., Benetti, L., Alphen, P. V., & Cristia, A. (2023). Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences*, *120*(52), e2300671120. https://doi.org/10.1073/pnas.2300671120

Bialystok, E., Luk, G., Peets, K. F., & Sujin, Y. (2010). Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism: Language and Cognition*, *13*(4), 525–531. https://doi.org/10.1017/S1366728909990423

Byers-Heinlein, K. (2013). Parental language mixing: Its measurement and the relation of mixed input to young bilingual children's vocabulary size. *Bilingualism: Language and Cognition*, *16*(1), 32–48. https://doi.org/10.1017/S1366728912000120

Byers-Heinlein, K. (2015). Methods for studying infant bilingualism. In J. W. Schwieter (Ed.), *The Cambridge handbook of bilingual processing* (pp. 133–154). Cambridge University Press.

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a Tseltal Mayan village. *Child Development*, *91*(5), 1819–1835. https://doi.org/10.1111/cdev.13349

Casillas, M., Brown, P., & Levinson, S. C. (2021). Early language experience in a Papuan community. *Journal of Child Language*, *48*(4), 792–814. https://doi.org/10.1017/S0305000920000549

Cristia, A. (2023). A systematic review suggests marked differences in the prevalence of infant‐directed vocalization across groups of populations. *Developmental Science*, *26*(1), e13265. https://doi.org/10.1111/desc.13265

Cychosz, M., Cristia, A., Bergelson, E., Casillas, M., Baudet, G., Warlaumont, A. S., Scaff, C., Yankowitz, L., & Seidl, A. (2021). Vocal development in a large‐scale crosslinguistic corpus. *Developmental Science*, *24*(5), e13090. https://doi.org/10.1111/desc.13090

de Boysson-Bardies, B., Hallé, P., Sagart, L., & Durand, C. (1989). A crosslinguistic investigation of vowel formants in babbling. *Journal of Child Language*, *16*(1), 1–17. https://doi.org/10.1017/S0305000900013404

de Boysson-Bardies, B., & Vihman, M. M. (1991). Adaptation to language: Evidence from babbling and first words in four languages. *Language*, *67*(2), 297–319. https://doi.org/10.1353/lan.1991.0045

Erickson, L. C., & Newman, R. S. (2017). Influences of background noise on infants and children. *Current Directions in Psychological Science*, *26*(5), 451–457. https://doi.org/10.1177/0963721417709087

Fibla, L., Sebastian-Galles, N., & Cristia, A. (2022). Is there a bilingual disadvantage for word segmentation? A computational modeling approach. *Journal of Child Language*, *49*(6), 1119–1146. https://doi.org/10.1017/S0305000921000568

Fox, J., & Weisberg, S. (2019). An R companion to applied regression (R package version 3.0-6) [Computer software]. https://CRAN.R-project.org/package=car

Goffman, L. (1999). Prosodic influences on speech production in children with specific language impairment and speech deficits: kinematic, acoustic, and transcription evidence. *Journal of Speech, Language, and Hearing Research*, *42*(6), 1499–1517. https://doi.org/10.1044/jslhr.4206.1499

Hitczenko, K., Bergelson, E., Casillas, M., Colleran, H., Cychosz, M., Grosjean, P., Hamrick, L. R., Kelleher, B. L., Scaff, C., & Seidl, A. (2023). The development of canonical proportion continues past toddlerhood. *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*, 1210–1214. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2023/full_papers/774.pdf

Hitczenko, K., Peurey, L., Harvard, W. N., Tey, K. J., Seidl, A., Semenzin, C., Scaff, C., Lavechin, M., Kelleher, B., Hamrick, L., Gautheron, L., Cychosz, M., Casillas, M., & Cristia, A. (submitted). Speech maturity dataset: A cross-cultural corpus of naturalistic child and adult vocalizations.

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of Child Language*, *39*(1), 1–27. https://doi.org/10.1017/S0305000910000759

Jung, J., & Houston, D. (2020). The relationship between the onset of canonical syllables and speech perception skills in children with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *63*(2), 393–404. https://doi.org/10.1044/2019_JSLHR-19-00158

Lambrecht Smith, S., Roberts, J. A., Locke, J. L., & Tozer, R. (2010). An exploratory study of the development of early syllable structure in reading-impaired children.

Journal of Learning Disabilities, *43*(4), 294–307. https://doi.org/10.1177/0022219410369094

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. *Interspeech*. https://doi.org/10.21437/Interspeech.2020-1690

Lee, C.-C., Jhang, Y., Relyea, G., Chen, L., & Oller, D. K. (2018). Babbling development as seen in canonical babbling ratios: A naturalistic evaluation of all-day recordings. *Infant Behavior and Development*, *50*, 140–153. https://doi.org/10.1016/j.infbeh.2017.12.002

Levitt, A. G., & Wang, Q. (1991). Evidence for language-specific rhythmic influences in the reduplicative babbling of French-and English-learning infants. *Language and Speech*, *34*(3), 235–249. https://doi.org/10.1177/002383099103400302

Lieberman, M., Hagberg, B., Lohmander, A., & Miniscalco, C. (2024). Follow-up of expressive language and general development at 12, 18 and 36 months for children with no canonical babbling at 10 months. *Clinical Linguistics & Phonetics*, 1–15. https://doi.org/10.1080/02699206.2024.2418127

Ma, Y., Jonsson, L., Feng, T., Weisberg, T., Shao, T., Yao, Z., Zhang, D., Dill, S.-E., Guo, Y., & Zhang, Y. (2021). Variations in the home language environment and early language development in rural China. *International Journal of Environmental Research and Public Health*, *18*(5), 2671. https://doi.org/10.3390/ijerph18052671

Ma, Y., Zhang, X., Pappas, L., Rule, A., Gao, Y., Dill, S., Feng, T., Zhang, Y., Wang, H., Cunha, F., & Rozelle, S. (2024). Associations between urbanization and the home language environment: Evidence from a LENA study in rural and peri‐urban China. *Child Development*, *95*(2). https://doi.org/10.1111/cdev.14034

Maddieson, I. (2013). Syllable structure. In M. S. Dryer & M. Haspelmath (Eds.), WALS Online (v2020.4) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13950591 (Available online at http://wals.info/chapter/12, Accessed on 2025-03-08.)

Nathani, S., Oller, D. K., & Neal, A. R. (2007). On the robustness of vocal development: An examination of infants with moderate-to-severe hearing loss and additional risk factors. *Journal of Speech, Language, and Hearing Research*, *50*(6), 1425–1444. https://doi.org/10.1044/1092-4388(2007/099)

Nip, I. S. B., & Green, J. R. (2013). Increases in cognitive and linguistic processing primarily account for increases in speaking rate with age. *Child Development*, *84*(4), 1324–1337. https://doi.org/10.1111/cdev.12052

Oller, D. K., & Eilers, R. E. (1988). The role of audition in infant babbling. *Child development*, 441-449. https://doi.org/10.2307/1130323

Oller, D. K., Eilers, R. E., Neal, A. R., & Cobo-Lewis, A. B. (1998). Late onset canonical babbling: A possible early marker of abnormal development. *American Journal on Mental Retardation*, *103*(3), 249–263. https://doi.org/10.1352/0895-8017(1998)103<0249:LOCBAP>2.0.CO;2

Oller, D. K., Eilers, R. E., Neal, A. R., & Schwartz, H. K. (1999). Precursors to speech in infancy: The prediction of speech and language disorders. *Journal of Communication Disorders*, *32*(4), 223–245. https://doi.org/10.1016/s0021-9924(99)00013-1

Oller, D. K., Pearson, B. Z., & Cobo-Lewis, A. B. (2007). Profile effects in early bilingual language and literacy. *Applied Psycholinguistics*, *28*(2), 191–230. https://doi.org/10.1017/S0142716407070117

Ott & Cychosz (under review). Can automated vocal analyses over child-centered audio recordings be used to predict speech-language development?

Patten, E., Belardi, K., Baranek, G. T., Watson, L. R., Labban, J. D., & Oller, D. K. (2014). Vocal patterns in infants with autism spectrum disorder: Canonical babbling status and vocalization frequency. *Journal of Autism and Developmental Disorders*, *44*, 2413–2428. https://doi.org/10.1007/s10803-014-2047-4

Paul, R., Fuerst, Y., Ramsay, G., Chawarska, K., & Klin, A. (2011). Out of the mouths of babies: Vocal production in infant siblings of children with ASD: Vocalizations in infant siblings. *Journal of Child Psychology and Psychiatry*, *52*(5), 588–598. https://doi.org/10.1111/j.1469-7610.2010.02332.x

Poulin-Dubois, D., & Goodz, N. (2001). Language differentiation in bilingual infants: Evidence from babbling. In J. Cenoz & F. Genesee (Eds.), *Trends in Language Acquisition Research* (Vol. 1, pp. 95–106). John Benjamins Publishing Company. https://doi.org/10.1075/tilar.1.06pou

R Core Team. (2013). R: A language and environment for statistical computing (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Semenzin, C., Hamrick, L., Seidl, A., Kelleher, B. L., & Cristia, A. (2021). Describing vocalizations in young children: A big data approach through citizen science annotation. *Journal of Speech, Language, and Hearing Research*, *64*(7), 2401–2416. https://doi.org/10.1044/2021_JSLHR-20-00661

Stringer, H. (2021). Phonological awareness: What comes before letters and sounds? Getting children ready for phonics [Handout]. Newcastle University. https://research.ncl.ac.uk/media/sites/researchwebsites/languageinterventionintheearlyyears/Lively_handout_HStringer.pdf

Sundara, M., Ward, N., Conboy, B., & Kuhl, P. K. (2020). Exposure to a second language in infancy alters speech production. *Bilingualism: Language and Cognition*, *23*(5), 978–991. https://doi.org/10.1017/S1366728919000853

Wickham, H. (2023). ggplot2: Create elegant data visualizations using the grammar of graphics (Version 3.4.0) [Computer software]. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/package=ggplot2

Vogt, P., Mastin, J. D., & Aussems, S. (2015). Early vocabulary development in rural and urban Mozambique. *Child Development Research*, *2015*, 1–15. https://doi.org/10.1155/2015/189195

Yankowitz, L. D., Petrulla, V., Plate, S., Tunc, B., Guthrie, W., Meera, S. S., Tena, K., Pandey, J., Swanson, M. R., Pruett, J. R., Cola, M., Russell, A., Marrus, N., Hazlett, H. C., Botteron, K., Constantino, J. N., Dager, S. R., Estes, A., Zwaigenbaum, L., … The IBIS Network. (2022). Infants later diagnosed with autism have lower canonical babbling ratios in the first year of life. *Molecular Autism*, *13*(1), 28. https://doi.org/10.1186/s13229-022-00503-8

Zheng, Z., Degotardi, S., Sweller, N., & Djonov, E. (2023). Effects of multilingualism on Australian infants' language environments in early childhood education centers. *Infant Behavior and Development*, *70*, 101799. https://doi.org/10.1016/j.infbeh.2022.101799

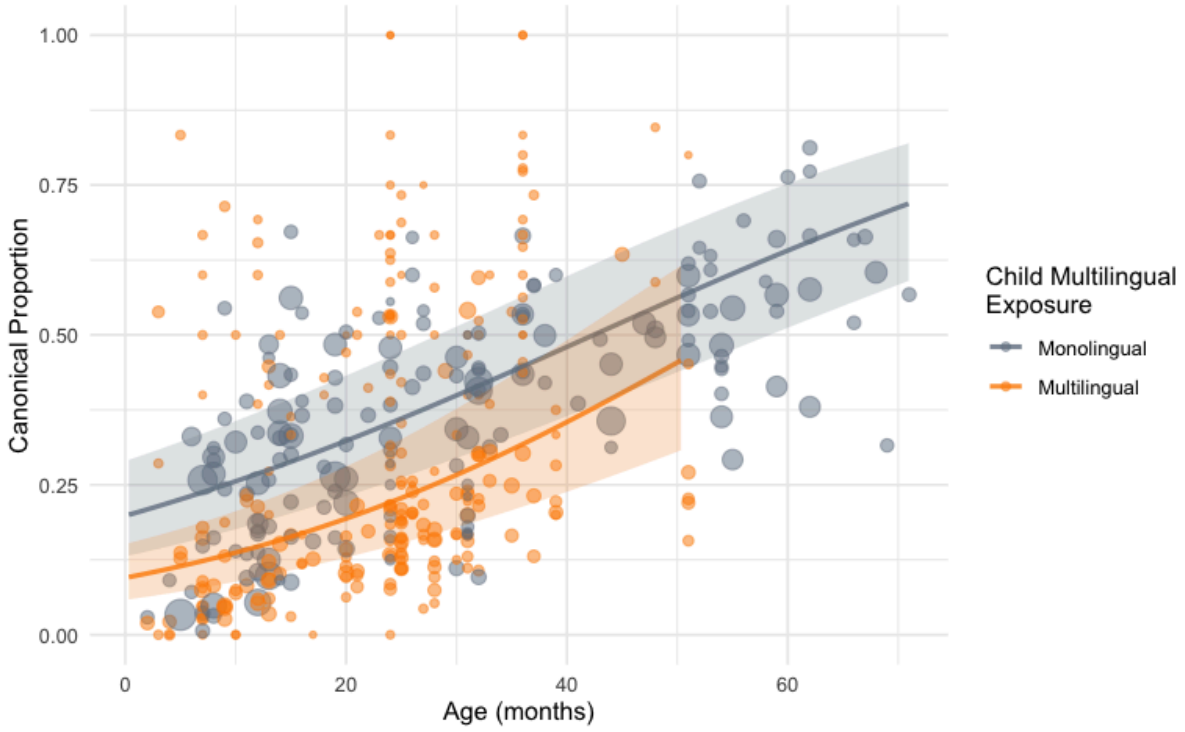
Supplementary Materials: The Development of Canonical Proportion as a Function of Community, Multilingualism, and Target Language's Syllable Complexity

Section 1

This section presents an additional analysis controlling for population imbalances in the dataset by balancing the number of children from Solomon Islands. In the original dataset, most multilingual children were from Solomon Islands, so a difference as a function of multilingualism may be between children from Solomon Islands versus elsewhere. To address this, we subset the Solomon Island corpus by randomly selecting only 40 children, matching the number of children from other corpora (see Table 1). This analysis included 213 children (110 monolinguals, 103 multilinguals) from the age range of 2-76 months old, examining CP as a function of multilingualism. The results revealed a significant main effect of multilingualism (Estimate=-.58, SE=.18, z=-3.22, p<.005), indicating that monolingual children had higher CP measures compared to multilingual children. Both age (Estimate=.49, SE=.06, z=8.69, p<.001) and quadratic age (Estimate=-.11, SE=.04, z=-2.62, p<.01) were also significant predictors of CP, suggesting that CP increase with age, but the growth slows with age. As in the main analysis, no significant interaction between age and multilingualism was observed. Supplementary Figure 1 shows the CP distribution in monolingual and multilingual children as a function of age.

Supplementary Figure 1

Canonical proportions by Age and Multilingual Exposure including only 40 children from the Solomon dataset



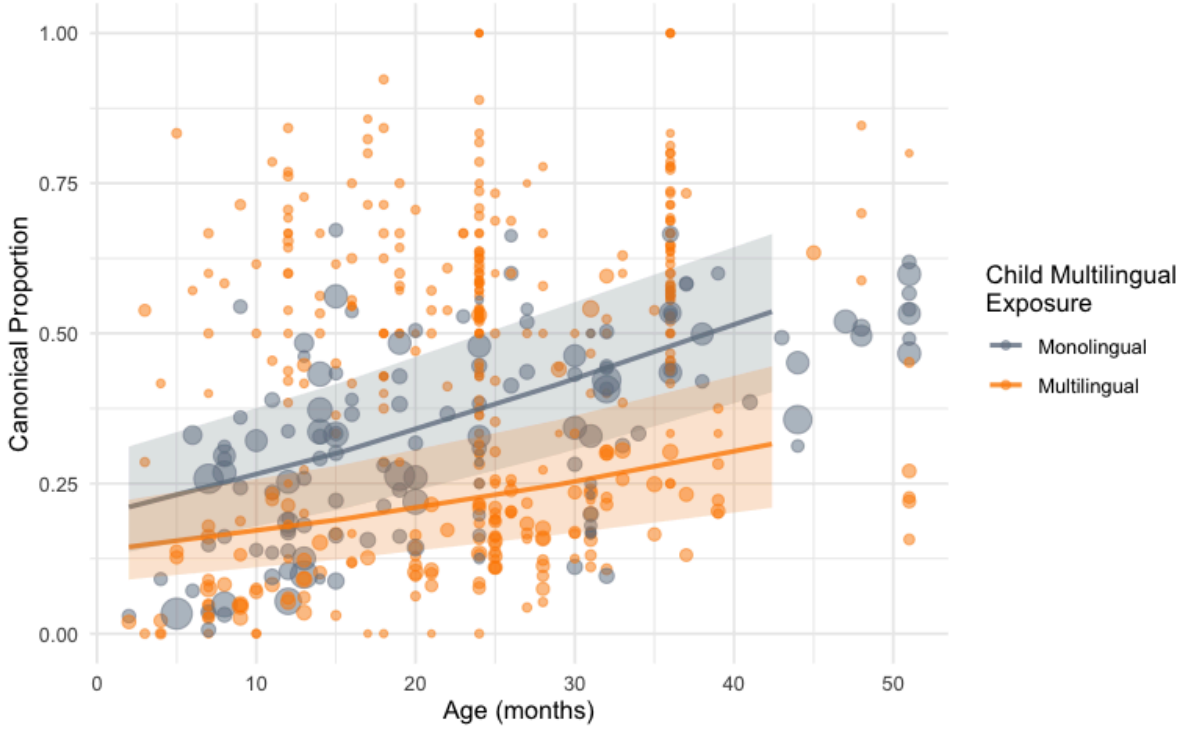
Note. The regression line represents the fitted model, and the shaded bands surrounding the line represent the 95% confidence intervals. Each data point represents a single child.

Section 2

This section gives an analysis subsetting to shared age ranges, examining CP as a function of multilingualism. For this analysis, we included 350 children (89 monolinguals, 261 multilinguals) from the age range of 2-51 months old. The results revealed a significant main effect of multilingualism (Estimate=-.74, SE=.18, z=-3.97, p<.001), indicating that monolingual children demonstrated higher CP measures compared to multilingual ones. Age was a significant predictor of CP (Estimate=.49, SE=.06, z=8.26, p<.001). No significant effect was found in quadratic terms for age. No significant interaction between age and multilingualism was also observed. Supplementary Figure 2 shows the CP distribution in monolingual and multilingual children across the children's age.

Supplementary Figure 2

Canonical proportions by Age and Multilingual Exposure subsetting to the shared age range



Note. The regression line represents the fitted model, and the shaded bands surrounding the line represent the 95% confidence intervals. Each data point represents a single child.

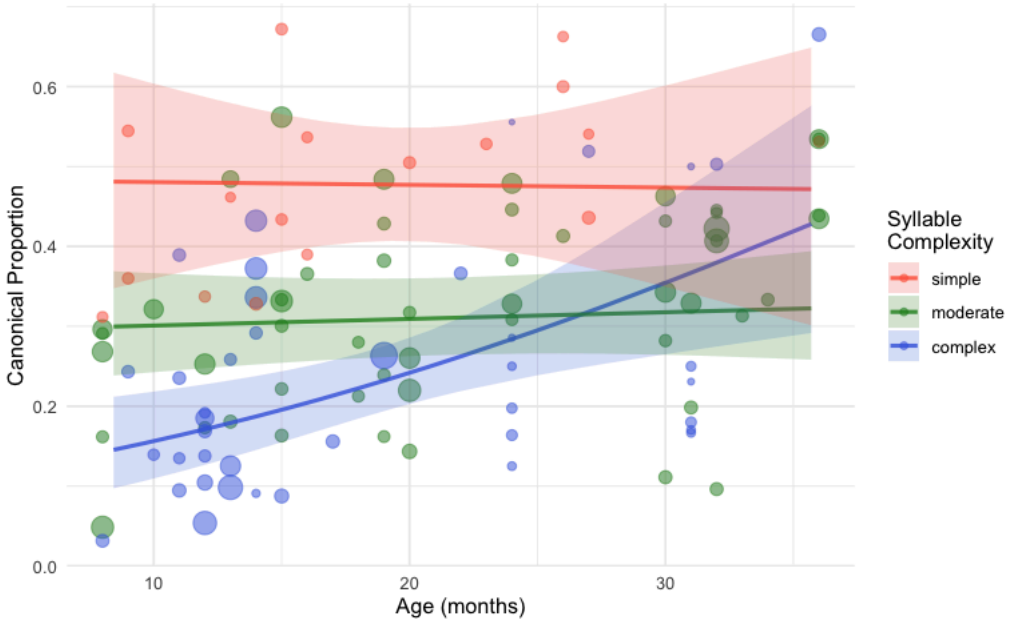
Section 3

This section presents an analysis restricted to the shared age range, examining the relationship between syllable complexity and CP among monolingual children only. To ensure comparability across groups, we included only 70 children (16 simple syllable complexity, 24 moderate, 30 complex) from 8 to 36 months old. The only significant effect was an interaction between syllable complexity and age. Supplementary Table 1 compares model estimates and standard errors between the primary analysis, which includes all monolingual children, and this subsetted analysis. From this table and Supplementary Figure 3, we conclude that the lack of significant differences between simple, moderate, and

complex is partly due to power (notice the larger SEs in the subset analysis) and partly due to different trajectories found within this age range (notice that simple and moderate do not exhibit the positive slope with age, in contrast with complex in this subset analysis, and with the data as a whole, in the analysis presented in the main text).

Supplementary Figure 3

Canonical proportions by Age and Syllable Complexity in Monolingual Children subsetting to the shared age range



Note. The regression line represents the fitted model, and the shaded bands surrounding the line represent 95% confidence intervals. Each data point represents a single child.

Supplementary Table 1

Comparison of Syllable Complexity Effects on CP Between Full and Subsetted Models

term	Estimate (Main)	SE (Main)	Estimate (Subset)	SE (Subset)
(Intercept)	0.08	0.12	-0.27	0.38
syl_commoderate	-0.70	0.15	-0.78	0.41
syl_comcomplex	-0.92	0.27	-0.35	0.49
month_age_z	0.36	0.12	-0.02	0.39
month_age_sq_z	-0.11	0.09	-0.74	0.69
syl_commoderate:month_age_z	-0.01	0.14	0.10	0.42
syl_comcomplex:month_age_z	0.46	0.30	1.02	0.51
syl_commoderate:month_age_sq_z	0.10	0.10	-0.52	0.76
syl_comcomplex:month_age_sq_z	-0.57	0.39	0.70	0.96

Note. This table compares model estimates and standard errors (SE) for syllable complexity and age effects on CP between the full dataset and the subsetted analysis (8-36 months). Estimates represent log-odds from a generalized linear mixed model with binomial family. SE = Standard Error.