**Demographic Biases in Naturalistic Language Recordings in the CHILDES Database**

Camila Scaff1,2, Georgia Loukatou3,, Alejandrina Cristia2, and Naomi Havron4

1 Institute of Evolutionary Medicine, University of Zurich

2 Laboratoire de Sciences Cognitives et de Psycholinguistique,

Département d’études cognitives, ENS, EHESS, CNRS, PSL University

3 Stanford University

4 Haifa University

**Abstract**

In recent years, the importance of estimating demographic biases in research has become apparent. Here we provide a systematic review of the CHILDES database, the major source of naturalistic recordings of children’s linguistic environment. We analyzed the database according to four dimensions considered central to language learning: SES, urbanization, family structure, and language. We present descriptive statistics of each dimension to assess whether naturalistic recordings were biased regarding the demographics of the countries and the families recorded within them. We find that CHILDES’s recordings overrepresented wealthier countries and higher parental education levels; urban settings; and smaller households. Middle- and higher-class participants were likewise over-represented. The corpora were not representative of their countries in terms of urbanization either - with a larger percentage of families residing in urban settings than is overall true for their respective countries. In terms of family structure, nuclear families were more prevalent than in the countries where the data was collected. Last, we found that corpora were linguistically diverse, but we estimate that these recordings underrepresent bilingual and multilingual households. We conclude that researchers should be careful when generalizing from naturalistic recordings of children's input and output obtained from CHILDES.

Since its foundation in 1984, CHILDES (the Child Language Data Exchange System, MacWhinney, 2000) has been the major source of naturalistic recordings and transcript data for researchers studying language acquisition. Moreover, it has contributed to establishing seminal work in various domains such as Theory of Mind (Bartsch & Wellman, 1995) and memory (Anderson & Schooler, 1991). It has also inspired theories that help understand the connection between language input and development (Christiansen et al., 1998). This paper aims to answer the question: Do CHILDES' naturalistic recordings, containing information about both input and output, come from a representative sample of the world's populations? If the answer is no, the generalizability of related findings may be limited, necessitating cautious interpretation.

Researchers in many fields are increasingly aware of the bias towards WEIRD populations (i.e., Western, Educated, Industrial, Rich, Democratic; Henrich 2010) in the samples they study (Cychosz & Cristia, 2022; Nielsen et al., 2017; Moriguchi, 2021; Singh et al., 2023). Recent calls have been made to diversify research in psychology, cognitive, and developmental science to address this issue (Majid & Levinson, 2010; Kidd & Garcia, 2022; Blasi et al. 2022). For instance, Kidd and Garcia (2022) systematically reviewed publications from child-language journals. They revealed biases towards specific continents (North America and Europe) and languages (English, Spanish, French), highlighting the need for increased diversity in the populations and languages studied.  
 Every area of research conceptualizes different dimensions relevant to explaining variance for a given research topic. In what follows we present the four most relevant dimensions thought to play a role in language acquisition, particularly related to naturalistic recordings: socioeconomic status, urbanization, family structure, and language.

#### Socioeconomic Status

It is beyond the scope of this paper to detail all the theories that attempt to account for the complex causal pathways that may connect socio-economic status (SES) to children's language environments. We recommend Rowe (2018) as a starting point for readers interested in this literature, along with Golinkoff et al. (2019) and Sperry et al. (2019) for diverse theoretical perspectives. In this paper, we do not take a stance on whether, how, or why SES and language environments relate to each other, but merely indicate that SES is one of the factors that has been repeatedly studied in the context of early language acquisition, particularly related to input. For example, Hoff (2003a) compared the speech of high- versus low-SES American mothers. College-educated mothers produced more utterances to their children, with more diverse vocabulary, longer phrases, and a higher number of utterances continuing a topic the child had brought up. Similar findings can be seen in other studies (Hart & Risley, 1995; Hoff-Ginsberg 1990; Hoff, 2003b; Huttenlocher, et al., 2002, 2007; see also Dailey & Bergelson, 2022; [Piot et al., 2021](https://www.zotero.org/google-docs/?9vsogx); for meta-analyses supporting the link; and Bergelson et al., 2023 for a large-scale study finding no significant SES effects). In the developmental literature, SES is primarily indexed by parents’ education (Ensminger et al., 2003; Hoff, 2003a), but can also be indexed by parental income, occupation, or a composite measure of these three (e.g. using the Hollingshead index; Hollingshead, 1975).

**Urbanization**

This dimension represents differences across populations in terms of the extent to which they are organized around urban sites, which often entail differences in job opportunities, living conditions, and infrastructure. Within the general theoretical framework of language socialization, there have been proposals that societies varying in their urbanization have differing views and values about the role of children in conversations (e.g., Sharma & Levine, 1998; Richman et al., 1992; Draper & Harpending, 1987; Keller, 2012). For instance, Keller (2012) discusses three prototypical cases, urban, rural, and hybrid, which differ in terms of their goals for children: Urban families aim for child psychological independence, rural families for child physical autonomy and interdependence, and hybrid families for some mix. Vogt et al. (2020) employed this classification to interpret their results on multimodal language use across three samples: urban Dutch, urban Mozambique, and rural Mozambique. They found that the number of gestures, gesture-speech alignment, and gesture types all vary across the three groups in ways that can be related to Keller's typology. Similarly, Cristia's (2023) systematic review uses urbanization status to describe the frequency of child-directed vocalization, finding that it is far less frequent in rural than urban communities.

**Family Structure**In this paper, we use the concept of family structure to encompass a variety of related aspects such as birth order and the way that caregiving is structured, each of which has an impact on child and language development (e.g., Blake, 1981; Duncan & Paradis, 2020; Havron et al., 2019, 2022, Hoff, 1993; Tomasello et al., 1995; Bornstein et al., 2019). For instance, children with older siblings exhibit lower language skills compared to first-born children across various cultures (e.g., Pyere et al., 2016 in France; Havron et al., 2022 in Singapore; [Zambrana](https://scholar.google.co.il/citations?user=a802HlgAAAAJ&hl=en&oi=sra), et al., 2012 in Norway), although second-born children might benefit from overhearing conversations between their caregivers and older sibling (Oshima-Takane et al., 1996). As for caregiving structures, in many middle-class Euro-American families, parents typically assume primary responsibility for children, with the mother as the primary caregiver (e.g. Bakermans-Kranenburg et al., 2004; Huttenlocher et al., 2010; Ispa et al., 2004; Pan et al., 2005). However, certain cultures, like the Turkish families described by Isleyen (2021), adopt a different approach, with nuclear families living in separate apartments but sharing common spaces and caregiving responsibilities, resulting in extensive support networks.

#### 

#### Languages

Characterizing the diversity of participant samples in terms of language (Blasi et al, 2022; Kidd & Garcia, 2022) is an important factor in language acquisition, allowing us to explore how *purely* linguistic factors shape and influence the development of language skills in children. For example, based on transcriptions of conversations, Pye and colleagues documented that K’iche' Mayan children frequently use and understand passive constructions from a very young age, unlike their English-speaking peers, refuting the idea that passive constructions can only emerge later in development (Pye, 1980; Pye & Poz, 1988).

Similarly, many Indo-European languages show a strong noun bias in early vocabularies (a bias for acquiring words for concrete referential objects rather than actions), which has been claimed to be a universal feature of early language acquisition. However, studies have shown that in some Mayan languages, including Tseltal and Tsotsil (Casillas et al., 2024; De Leon, 1999), there is little to no evidence for a noun bias, and argue for a verb bias instead. This highlights significant cross-linguistic variation (Casillas et al., 2024).  
 Another major dimension and sub-field in developmental science is the study of bilingualism or multilingualism (for an introductory review see Höhle et al., 2019, and McCabe et al., 2013). There is evidence that monolingual and bilingual early language development differs in some aspects, particularly regarding phonological acquisition and word learning. For example, monolingual infants' ability to discriminate non-native sounds declines during the first year of age, whereas infants exposed to more language(s) maintain the discrimination window for a longer period (Singh et al., 2022). In terms of input, bilingual linguistic exposure is divided between two or more native languages. It has been shown that the amount of exposure to each native language can affect bilingual infants' speech discrimination abilities (Garcia-Sierra et al., 2011).

### **The current study**

Inspired by calls to document the demographic and linguistic diversity of participant samples (Kidd & Garcia, 2021; Singh et al, 2023), we sought to provide a systematic analysis of the naturalistic speech corpora of the CHILDES database by quantifying the diversity of each dimension introduced above (SES, urbanization, family structure, and languages).

# **Method**

#### Inclusion criteria

We extracted information from the online CHILDES database in 2021, at which point there were 430 corpora. We excluded corpora on clinical populations (N = 69), in-lab recordings (N = 26), and elicitation tasks (N = 114), because they would have introduced additional sources of variation and biases. For example, the inclusion of clinical populations may highlight differences in the prevalence and access to diagnosis in different countries. Given our interest in naturalistic input and output represented in CHILDES, we also excluded corpora where only child speech was transcribed (N = 19) and diary studies (N = 11). Finally, we excluded 11 corpora that were not available, resulting in a final sample of 180 corpora containing naturalistic recordings of children's input and output.

#### Demographic Data Collection

Data on participant demographics was extracted from the references provided for each corpus in CHILDES such as articles, book chapters, dissertations, and the actual content of transcribed recordings. We contacted corpus curators for missing information, and over a third (103 corpora, 32%) provided additional data or confirmed missing information. Unfortunately, curators for 50 corpora (15%) could not be contacted due to unresponsive email addresses (38 corpora), or the curator’s passing (12 corpora).

#### Analyses

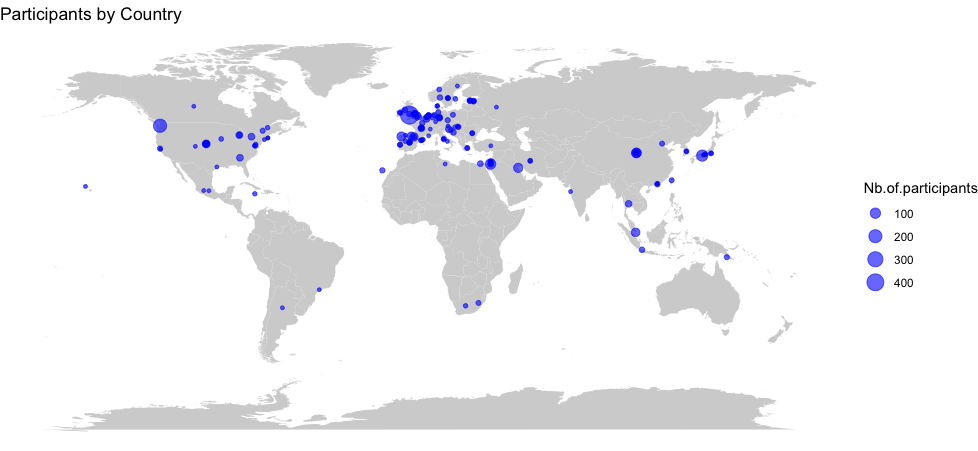
Analyses and visualizations were carried out using R (version 4.1.2, R Core Team, 2020) and ggplot2 (Wickham, 2016). Data, scripts, and Supplementary Materials are available on the OSF (Scaff et al., 2024). Following Ghai's (2022) recommendation to enhance the description of diversity in behavioral sciences, we examined different levels, from a macro-level, involving broad country comparisons, to a corpus-level, where we analyzed individual corpora for a more detailed and nuanced understanding of the data.

## 

## Results

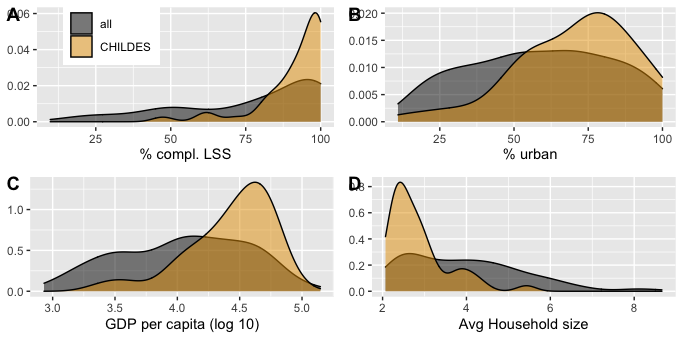
## Do Countries in Our CHILDES Sample Represent the World's Countries?

The 180 corpora included in this study represent 48 different countries or territories across all populated continents (see Figure 1), although 28 countries, representing 151 corpora (84%), belong to the OECD (Organisation for Economic Co-operation and Development). Meanwhile, OECD countries represent only 19.5% of the world’s countries. The country with the most corpora is the United States with 30 different corpora, followed by Spain with 24, followed by the United Kingdom and France with 11.



**Figure 1.** *Geographical Distribution of the Corpora in CHILDES.* Each circle represents a corpus, and its size is proportional to the number of participants.

We checked for representativeness given the four key dimensions described in the Introduction (see Figure 2), by using public information on the represented countries (see SM1 for details). Countries in our sample of CHILDES differed significantly from the distribution of countries in the world according to unpaired tests that did not assume equality of variances (Welch's *t*): a higher proportion of the population completed lower secondary school than the worldwide sample (% compl. LSS, *t*(130.41) = -6.19, *p* < .001); the countries were more urban (% urban, *t*(79.48) = -3.44, *p* < .001) and richer (log GDP per capita, *t*(102.35) = -6.02, *p* < .001); and they had smaller households (average household size, *t*(118.8) = 7.12, *p* < .001).



**Figure 2.** *Comparison of Country-Level Descriptors Between World Countries and Our CHILDES Sample.* Density plots show the distribution of country-level descriptors for all world countries (gray) versus countries represented in our sample of CHILDES (orange). % complete. LSS is the percentage of the country's population completing lower secondary school. % urban is the percent of the country's population residing in urban (as opposed to rural) locations. GDP stands for Gross Domestic Product. Avg Household size stands for average household size. Data from the World Bank, Our World in Data, and the United Nations. See SM1 for additional details.

#### How Representative and Diverse are Participant Samples?

We next investigated variables within each corpus, as far as possible given substantial missing data (see Table 1 and figures below regarding the number of corpora represented in each analysis). The information collected gives us more detail about the characteristics of the individual families that compose our sample of CHILDES.

**Table 1.** *Variables of interest.*

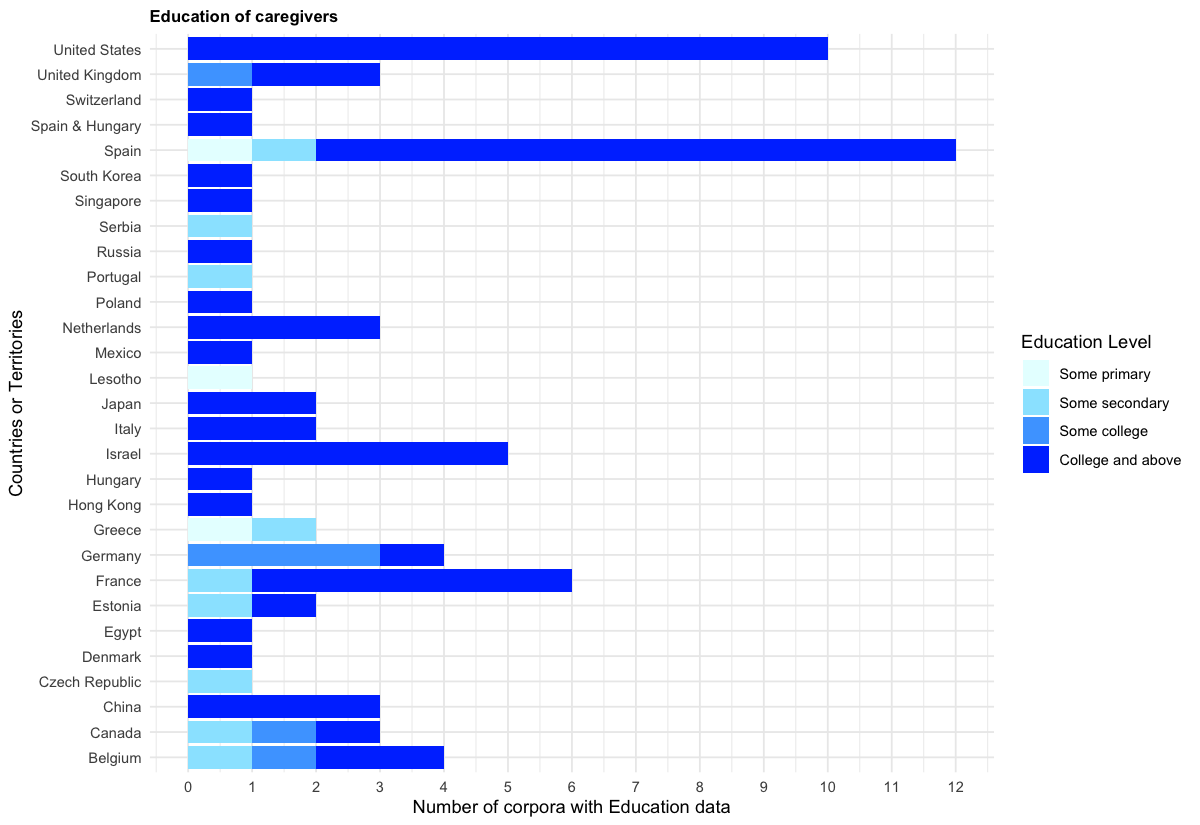
|  |  |  |  |
| --- | --- | --- | --- |
| **Dimension** | **Corpus level variable** | **Definition of levels** | **N corpus (% of corpora)** |
| SES | Parent's socio-economic status | Low, mid, and/or high | 102 (57%) |
| SES | Parents’ education level | Highest level of education completed: Primary, High school, Undergraduate, Masters, PhD | 80 (44%) |
| SES | Parents’ occupation | Parents’ activity or profession | 84 (47%) |
| Urbanization | Type of community | Urban, rural, mixed | 65 (36%) |
| Family structure | Household composition | Whether the family was composed primarily of caregivers and children (nuclear) or also included other family members: grandparents, cousins, uncles, and aunts (extended) | 69 (38%) |
| Family structure | Percent children with sibling(s) | The percentage of children in the corpus who had at least one sibling | 93 (52%) |
| Family structure | Average number of siblings | How many siblings’ children had on average (including children with zero siblings) | 83 (46%) |
| Language | Language(s) spoken | Which languages were spoken in the transcripts and by the family | 180 (100%) |
| Language | Lingual status | Whether more than one language is spoken in the corpus (regardless of how many languages are spoken in the family or the community at large) | 110 (61%) |

#### 

#### Socio-economic Status: SES, Education, and Occupation

Most corpora provided global SES information. Among these, the vast majority of families were of middle or higher SES. Although about a fifth of the corpora were described as a mixed sample spanning lower to middle or higher SES, a very small percentage were described as families of low SES. As we show in SM2, this general distribution over-represents higher SES families.

A smaller proportion of corpora (additionally) reported parental education level (Figure 3). Among them, three-quarters included at least some children whose parents had a graduate or postgraduate degree. In fact, over half of the corpora did not contain data from any children whose parents had not completed a graduate degree. Overall, more educated parents are overrepresented given the source countries, which is reinforced by the observations that parents had professions linked to graduate-level education in over half of corpora (see SM3).

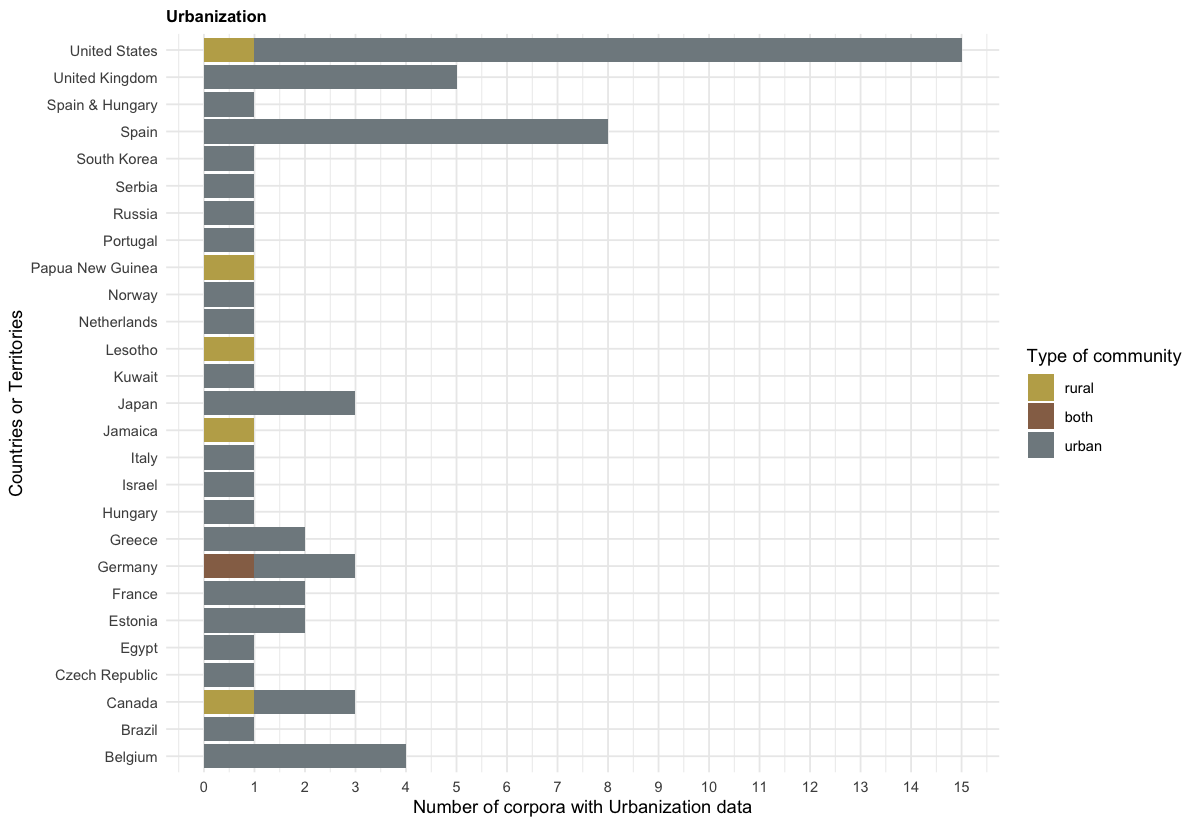


**Figure 3.** *Distribution of Minimum Levels of Parental Education by Country.*  "Spain & Hungary" corresponds to a corpus collected in two countries. The same applies for Figures 4-7.

#### 

#### Urbanization

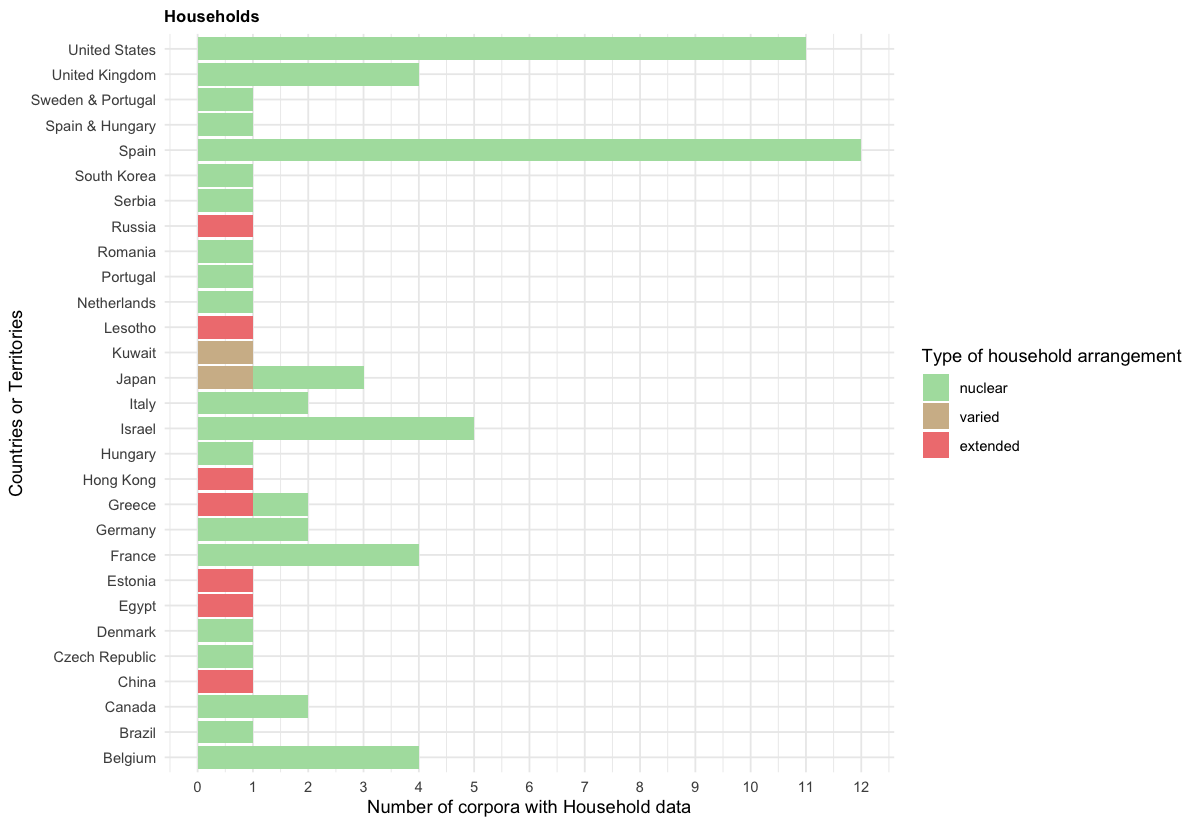
The majority of corpora were urban (Figure 4), with an over-representation of urban households given the source countries (see SM4).



**Figure 4.** *Distribution of Type of Community by Country.*

#### Family Structure: Household composition and number of siblings

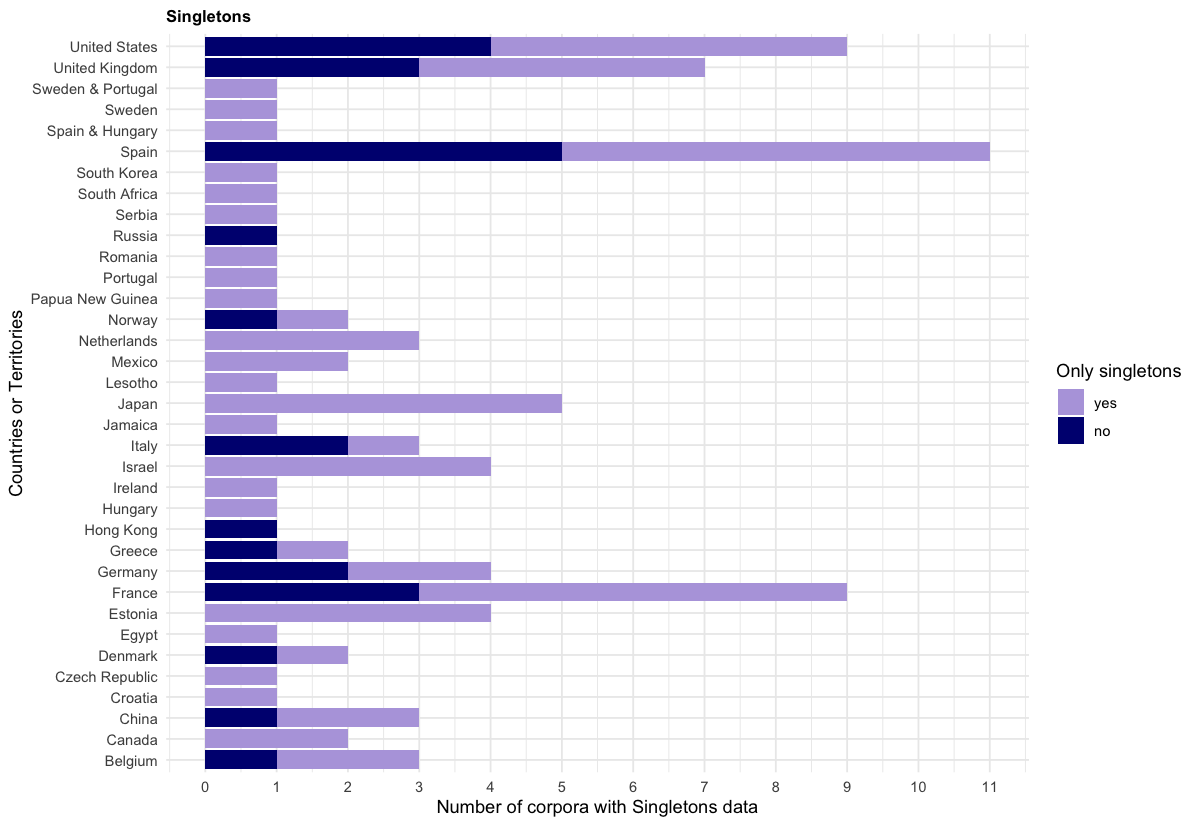
As Figure 5 shows, the overwhelming majority of corpora contained only nuclear families (i.e., parents and their children), and only a tiny percentage had a mix of nuclear and extended families. We were not able to find a systematic database allowing us to check whether this was representative of the source countries.



**Figure 5.** *Distribution of Household Arrangement by Country*. "Sweden & Portugal" corresponds to a corpus collected in two countries. The same applies for Figure 7.

#### About a quarter of the corpora in our CHILDES sample consist exclusively of singletons (children with no siblings), while the remaining include children who have at least one sibling (Figure 6). SM5 argues that this does not necessarily mean that singletons are underrepresented.

#### 

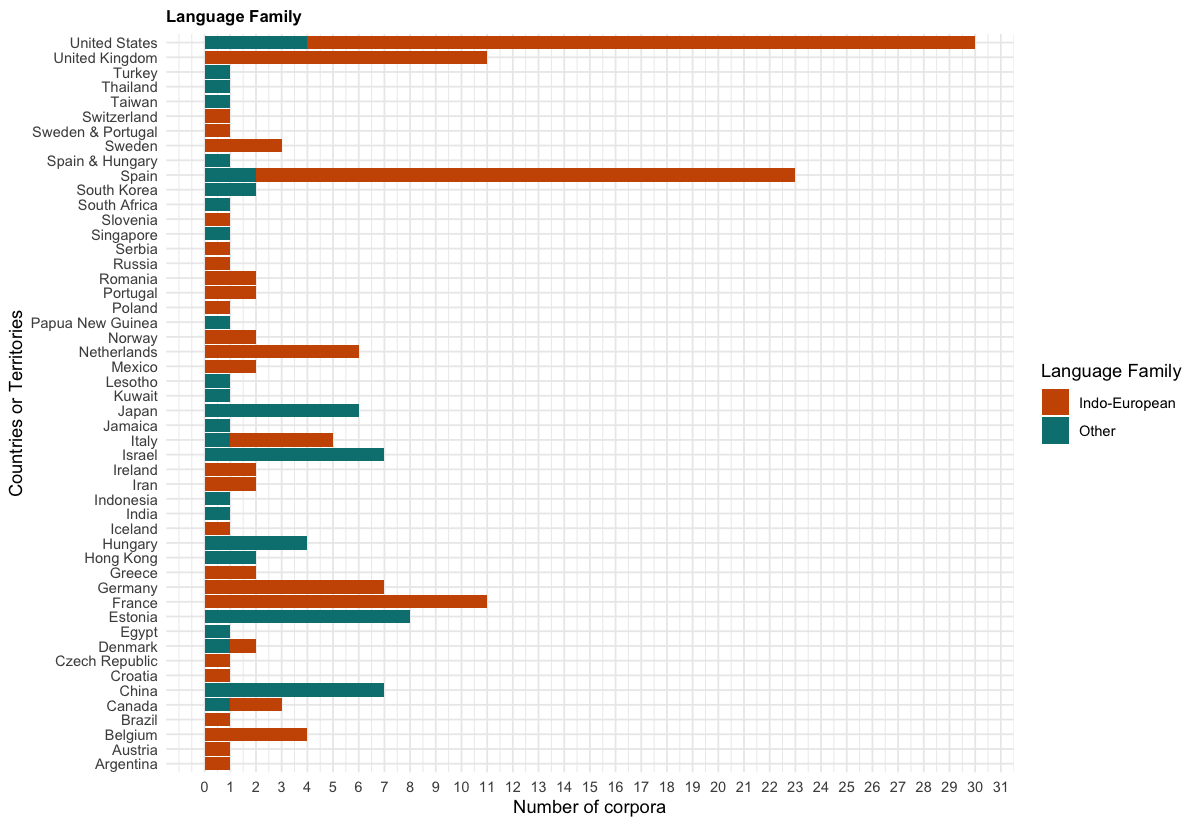


**Figure 6.** *Distribution of Singletons by Country.*

#### Languages

We find a total of 63 different languages or language combinations (for bilingual and multilingual children; see SM6). About a third of the included corpora were not monolingual. According to information in SM7, our sample of CHILDES underrepresents bilingual and multilingual households in the world.

As for linguistic diversity, most of the corpora pertain to Indo-European languages (Figure 7). English is the most prevalent language, both in terms of the number of children and the number of corpora, with 50 corpora (30 exclusively monolingual), comprising 28% of the included corpora. Additionally, it is the language most studied for bilingualism or multilingualism, with language combinations including Cantonese, Danish, Dutch, Farsi, French, Hebrew, Hungarian, Japanese, Mandarin, Portuguese, Russian, Spanish, and Swedish. The next most represented language in terms of different corpora is Spanish, with 25 corpora (14 monolingual).



**Figure 7.** *Distribution of Indo-European Languages by Country.*

# 

# **Discussion**

The current paper provides a systematic review of corpora relying on naturalistic recordings in the CHILDES database, a priceless public research resource. We assessed and characterized sampling bias by examining the database through four macro dimensions: SES, urbanization, family structure, and languages, all representing dimensions believed to be relevant for early language acquisition. We aimed to address the question of how biased the sample was in terms of these four dimensions when compared to the world's population, as well as when taking into account the countries in which the data were collected.

Thanks to the collective work of nearly 300 individuals (with deep thanks to the people acknowledged in SM8), we could include 180 corpora from 48 different countries, primarily OECD (84%). Despite a remarkable diversity in terms of languages and countries, analyses of the demographic characteristics of our CHILDES sample revealed several measures where the corpora do not represent the world population, nor did the corpora faithfully represent the countries in which data were collected, in terms of SES, urbanization, family structure, and linguistic dimensions.

In terms of SES, our macro-level analyses showed that represented countries had a higher percentage of individuals completing lower secondary school, and higher Gross Domestic Product (GDP) per capita, compared to the worldwide distribution. Our focused analyses further suggested that higher SES families with highly educated parents were more common than would have expected given the source countries. Considering the extensive literature linking SES and language acquisition, these are important biases to take into account.

The dataset's overrepresentation of highly educated parents extends beyond this educational/income bias. Notably, approximately half of the corpora feature parents engaged in research-related professions, with academics frequently studying their own (grand)children. In comparison, in the US general population, in 2020, only 6% of American citizens would fall under such definitions. The danger of extrapolating from such a sample goes beyond issues previously discussed in the context of convenience (homogeneous or heterogeneous) samples (Bornstein et al., 2013), and may concern the very fact that parents who pursue a career in academia (and likely in language-related areas) could be particular in their linguistic behavior.

We also uncovered significant biases in the urbanization dimension. Regarding urbanization, 91% of corpora were urban, compared to 59% of the world population, with only 5 corpora exclusively rural. The latter seemed to be reasonable given urbanization statistics for the countries where data were collected. Thus, in this case, the bias pertains only to which countries are represented, and not how data are sampled within the country. One challenge for future work will involve teasing apart the myriads of ways in which urbanization may relate to language outcomes, including not only the parental goals mentioned in the Introduction (Keller, 2012) but also factors that may be confounded with urbanization. For example, families in rural conditions may have less access to infrastructure, but they may be protected from certain types of stress. In fact, a study using long-form recordings suggested that infants growing up in formerly rural families that had migrated to urban centers were afforded less spoken input and had smaller vocabularies than infants whose families had remained rural (Ma et al., 2024).

In terms of family structure, the represented countries had smaller households compared to the world average. When comparing household composition in our CHILDES sample with the countries where data had been collected, we found mismatches, with likely a trend for underrepresenting extended families. A bias towards nuclear families limits our understanding of diverse caregiving structures, particularly alloparental involvement, in language acquisition. While multi-child families are well-represented in our CHILDES sample, the absence of siblings' speech in some transcripts (e.g., Loukatou et al., 2022) hinders insights into typical daily language interactions. In fact, even when recorded children reportedly had siblings, this does not mean that the siblings were recorded, transcribed, or even present during the recording session. Limited data on birth order also restricts robust conclusions on sibling composition in CHILDES and underscores the need for broader exploration of non-parental caregiving roles typically overlooked in studies focusing on predominantly WEIRD households.

Finally, we found that English and more broadly Indo-European languages were overrepresented in our CHILDES sample, a point already made by others (e.g., Christiansen et al., 2022). This aligns with trends in child development literature (Kidd & Garcia, 2022; Singh et al., 2023). Additionally, although a third of the corpora were not monolingual, this nonetheless underrepresents multilingualism in the world's population (Grosjean, 2024).

Although our CHILDES sample is more homogeneous than expected given global and national variations, there remains substantial diversity across all the dimensions we studied. Could researchers use this variation to study the association between, e.g., urbanization and language characteristics in such spontaneous, naturalistic samples? While excited at the prospect, we also fear that such an exploration would be fraught in terms of causal reading given how confounded the dimensions are in these CHILDES corpora. Put simply, since most samples are from nuclear families with well-educated parents living in urban sites, it is impossible to tease apart these three dimensions. Additionally, given that few corpora represent some combinations, it will be hard to understand the intersectionality of such factors in naturalistic input samples.

While making some strides in documenting the demographic characteristics of the people represented in naturalistic corpora in CHILDES, our work has several limitations we would like to acknowledge. First, we focused on four dimensions (SES, urbanization, family structure, and language), but others may well be relevant for language acquisition research. We hope that our sharing of materials will facilitate others' addition of novel descriptors. In particular, the World Values Survey contains information about several dimensions that could be relevant to language acquisition such as social norms and attitudes towards gender roles (World Values Survey Association, 2020).

Second, the absence of comprehensive data for certain variables restricted a thorough analysis of representativeness on some dimensions. This should not be seen as the fault of the corpus creator or curator, as these data were not always available to them or were not considered relevant at the time of data collection. After all, it is only recently that researchers in our field started sounding the alarm regarding underreporting of relevant demographic characteristics (e.g., Singh, et al., 2023; Singh et al., 2023).

In addition, as computational modeling becomes more prevalent, it is important to keep in mind what are the generalizations that modeling based on such data might arrive at, and how they represent the overarching mechanisms of language acquisition. A relevant example is the BabyLM Challenge, created with the aim of fostering consideration of developmental plausibility when developing language-related machine learning tools (Warstadt et al., 2023). We recommend that researchers re-using CHILDES' input and output data inform themselves about the characteristics of the people recorded and acknowledge potential limitations in demographic representation.

To conclude, we found that data from many countries and languages are represented in CHILDES, but that these countries did not constitute a representative sample of the world's countries, and that families were not representative of their own countries. We also noted low variability along certain dimensions (saliently SES and family structure) and the predominance of Indo-European languages, with English having the most recorded participants. Our discussion leads us to argue for the systematic inclusion of certain descriptors in speaker-level and corpus-level metadata, to track and quantify both diversity and representativity.

**Acknowledgements**

We are grateful to Lara Oliel and Clara Daoud for help entering corpus-level data; Sara Pisani for organizing communication with corpus curators; Brian MacWhinney for support and encouragement when retrieving information from corpus curators. We particularly want to thank the literally hundreds of researchers who contributed data to CHILDES, without which this project would be unthinkable; and the 103 curators who replied to our request.

AC acknowledges financial and institutional support from Agence Nationale de la Recherche (ANR-16-DATA-0004 ACLEW, ANR-17-EURE-0017, ANR-14-CE30-0003 MechELex); the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award; and the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (ExELang, Grant agreement No. 101001095). GL acknowledges financial support from Fondation Fyssen.

**Conflict of interest statement**

None of the authors have a conflict of interest to disclose.

# 

# **References**

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408. <https://doi.org/10.1111/j.1467-9280.1991.tb00174.x>

Scaff, C., Loukatou, G., Cristia, A., & Havron, N. (2024). Supplementary materials <https://osf.io/q9w82/?view_only=a013f1b25b8c4556b8248f12870402c9>

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. New York, NY: Oxford University Press.

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramirez-Esparza, N., Hamrick, L. R., & others. (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>.

Blake, J. (1981). Family size and the quality of children. *Demography*, *18*(4), 421–442. <https://doi.org/10.2307/2060941>

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.007

Bornstein, M. H., Jager, J., & Putnick, D. L. (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental Review*, *33*(4), 357–370. https://doi.org/10.1016/j.dr.2013.08.003

Bornstein, M. H., Putnick, D. L., & Suwalsky, J. T. (2019). Mother–infant interactions with firstborns and secondborns: A within-family study of European Americans. *Infant Behavior and Development*, *55*, 100–111. https://doi.org/10.1016/j.infbeh.2019.03.002

Casillas, M., Foushee, R., Méndez Girón, J., Polian, G., & Brown, P. (2024). Little evidence for a noun bias in Tseltal spontaneous speech. *First Language*, 01427237231216571. https://doi.org/10.1177/01427237231216571

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*(2-3), 221–268. https://doi.org/10.1080/016909698386528

Christiansen, M. H., Contreras Kallens, P., & Trecca, F. (2022). Toward a comparative approach to language acquisition. *Current Directions in Psychological Science*, *31*(2), 131–138. https://doi.org/10.1177/09637214211064692

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. (2022). Using big data from long-form recordings to study development and optimize societal impact. In R. Kail (Ed.), *Advances in Child Development and Behavior* (Vol. 62, pp. 1–36). Academic Press.

Dailey, S., & Bergelson, E. (2022). Language input to infants of different socioeconomic statuses: A quantitative meta-analysis. *Developmental Science*, *25*(3), e13192.

De León, L. (1999). Verbs in Tzotzil (Mayan) early syntactic development. *International Journal of Bilingualism*, *3*(2-3), 219–239.

Demuth, K. (2022). The acquisition of Sesotho. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition* (pp. 557–638). Psychology Press.

Draper, P., & Harpending, H. (2017). Parent investment and the child’s environment. In J. B. Lancaster, A. Rossi, & J. Altmann (Eds.), *Parenting across the life span* (pp. 207–236). Routledge.

Duncan, T. S., & Paradis, J. (2020). Home language environment and children’s second language acquisition: The special status of input from older siblings. *Journal of Child Language*, *47*(5), 982–1005. https://doi.org/10.1017/S0305000919000905

Ensminger, M. E., & Fothergill, K. E. (2014). A decade of measuring SES: What it tells us and where to go from here. In M. H. Bornstein & R. H. Bradley (Eds.), *Socioeconomic status, parenting, and child development* (pp. 13–27). Routledge.

Garcia-Sierra, A., Rivera-Gaxiola, M., Percaccio, C. R., Conboy, B. T., Romo, H., Klarman, L., … Kuhl, P. K. (2011). Bilingual language learning: An ERP study relating early brain responses to speech, language input, and later word production. *Journal of Phonetics*, *39*(4), 546–557. https://doi.org/10.1016/j.wocn.2011.07.002

Ghai, S. (2021). It’s time to reimagine sample diversity and retire the WEIRD dichotomy. *Nature Human Behaviour*, *5*(8), 971–972. https://doi.org/10.1038/s41562-021-01175-9

Golinkoff, R. M., Hoff, E., Rowe, M. L., Tamis-LeMonda, C. S., & Hirsh-Pasek, K. (2019). Language matters: Denying the existence of the 30-million-word gap has serious consequences. *Child Development*, *90*(3), 985–992. https://doi.org/10.1111/cdev.13128

Grosjean, F. (2024). The statistics of bilingualism. In J. F. Kroll & A. M. B. de Groot (Eds.), *On bilinguals and bilingualism* (pp. 138–147). Cambridge University Press.<https://doi.org/10.1017/9781009210409.010>

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: Paul H Brookes Publishing.

Havron, N., Lovcevic, I., Kee, M. Z., Chen, H., Chong, Y. S., Daniel, M., … Tsuji, S. (2022). The effect of older sibling, postnatal maternal stress, and household factors on language development in two-to four-year-old children. *Developmental Psychology*, *58*(11), 2096. https://doi.org/10.1037/dev0001395

Havron, N., Ramus, F., Heude, B., Forhan, A., Cristia, A., Peyre, H., & Group, E. M.-C. C. S. (2019). The effect of older siblings on language development as a function of age difference and sex. *Psychological Science*, *30*(9), 1333–1343. https://doi.org/10.1177/0956797619861436

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29. https://doi.org/10.1038/466029a

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*(5), 1368–1378. https://doi.org/10.1111/1467-8624.00612

Hoff, E. (2014). Causes and consequences of SES-related differences in parent-to-child speech. In M. H. Bornstein & R. H. Bradley (Eds.), *Socioeconomic status, parenting, and child development* (pp. 147–160). Routledge.

Hoff-Ginsberg, E. (1990). Maternal speech and the child’s development of syntax: A further look. *Journal of Child Language*, *17*(1), 85–99. https://doi.org/10.1017/S0305000900013091

Hoff-Ginsberg, E., & Krueger, W. M. (1991). Older siblings as conversational partners. *Merrill-Palmer Quarterly*, *37*(3), 465–482.

Höhle, B., Bijeljac-Babic, R., & Nazzi, T. (2020). Variability and stability in early language acquisition: Comparing monolingual and bilingual infants’ speech perception and word recognition. *Bilingualism: Language and Cognition*, *23*(1), 56–71. https://doi.org/10.1017/S1366728919000422

Huttenlocher, J., Vasilyeva, M., Waterfall, H. R., Vevea, J. L., & Hedges, L. V. (2007). The varieties of speech to young children. *Developmental Psychology*, *43*(5), 1062. https://doi.org/10.1037/0012-1649.43.5.1062

Isleyen, M. A. (2021). *Marital functioning and parenting in extended family living arrangements: A qualitative study in family buildings* (Master’s thesis). Istanbul Bilgi University, Istanbul, Turkey. https://doi.org/10.31234/osf.io/gpvqr

Keller, H. (2012). Autonomy and relatedness revisited: Cultural manifestations of universal human needs. *Child Development Perspectives*, *6*(1), 12–18. https://doi.org/10.1111/j.1750-8606.2011.00208.x

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, *42*(6), 703–735. https://doi.org/10.1177/01427237221124710

Leonardo, P., Havron, N., & Cristia, A. (2022). Socioeconomic status correlates with measures of language environment analysis (LENA) system: A meta-analysis. *Journal of Child Language*, *49*(5), 1037–1051. https://doi.org/10.1017/S0305000922000314

Lieven, E., & Stoll, S. (2013). Early communicative development in two cultures: A comparison of the communicative environments of children from two cultures. *Human Development*, *56*(3), 178–206. https://doi.org/10.1159/000351073

Loukatou, G., Scaff, C., Demuth, K., Cristia, A., & Havron, N. (2022). Child-directed and overheard input from different speakers in two distinct cultures. *Journal of Child Language*, *49*(6), 1173--1192. https://doi.org/10.1017/S0305000922000387

Ma, Y., Zhang, X., Pappas, L., Rule, A., Gao, Y., Dill, S. E., ... & 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), e74-e92. https://doi.org/10.1111/cdev.13718

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. MIT Press.

MacWhinney, B. (2019). Understanding spoken language through TalkBank. *Behavior Research Methods*, *51*, 1919–1927. https://doi.org/10.3758/s13428-018-1174-5

Majid, A., & Levinson, S. C. (2010). Weird languages have misled us, too. *Behavioral and Brain Sciences*, *33*(2-3), 103–103.<https://doi.org/10.1017/s0140525x1000018x>

McCabe, A., Tamis-LeMonda, C. S., Bornstein, M. H., Brockmeyer Cates, C., Golinkoff, R., Wishard Guerra, A., & others. (2013). Multilingual children beyond myths and toward best practices. *Social policy report,* (27)4. Society for Research in Child Development.

Moriguchi, Y. (2022). Beyond bias to Western participants, authors, and editors in developmental science. *Infant and Child Development*, *31*(1), e2256. https://doi.org/10.1002/icd.2256

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017

Oshima-Takane, Y., Goodz, E., & Derevensky, J. L. (1996). Birth order effects on early language development: Do secondborn children learn from overheard speech? *Child Development*, *67*(2), 621–634. https://doi.org/10.2307/1131833

Peyre, H., Bernard, J. Y., Hoertel, N., Forhan, A., Charles, M.-A., De Agostini, M., et al.others. (2016). Differential effects of factors influencing cognitive development at the age of 5-to-6 years. *Cognitive Development*, *40*, 152–162. https://doi.org/10.1016/j.cogdev.2016.10.003

Pye, C. L. (1980). *The acquisition of grammatical morphemes in Quiche Mayan* [PhD Thesis]. University of Pittsburgh.

Pye, C., & Poz, P. Q. (1988). *P*recocious passives (and antipassives) in Quiche Mayan. *Journal of Child Language, 15*(1), 129–135. https://doi.org/10.1017/S0305000900013616

Richman, A. L., Miller, P. M., & LeVine, R. A. (1992). Cultural and educational variations in maternal responsiveness. *Developmental Psychology*, *28*(4), 614. https://doi.org/10.1037/0012-1649.28.4.614

Rowe, M. L. (2018). Understanding socioeconomic differences in parents’ speech to children. *Child Development Perspectives*, *12*(2), 122–127. https://doi.org/10.1111/cdep.12274

Sarvasy, H. (2017). *A grammar of Nungon: A papuan language of Northeast New Guinea* (Vol. 4). Brill.

Scaff, C., Casillas, M., Stieglitz, J., & Cristia, A. (2024). Characterization of children’s verbal input in a forager-farmer population using long-form audio recordings and diverse input definitions. *Infancy*, *29*(2), 196–215. https://doi.org/10.1111/infa.12534

Sharma, D., & LeVine, R. A. (1998). Child care in India: A comparative developmental view of infant social environments. *New Directions for Child and Adolescent Development*, *81*, 45–67. https://doi.org/10.1002/cd.23219988105

Simon, K. R., Merz, E. C., He, X., & Noble, K. G. (2022). Environmental noise, brain structure, and language development in children. *Brain and Language*, *229*, 105112. https://doi.org/10.1016/j.bandl.2022.105112

Singh, L., Rajendra, S. J., & Mazuka, R. (2022). Diversity and representation in studies of infant perceptual narrowing. Child Development Perspectives, 16(4), 191-199. <https://doi.org/10.1111/cdep.12450>

Singh, L., Barokova, M. D., Baumgartner, H. A., Lopera-Perez, D. C., Omane, P. O., Sheskin, M., ... & Frank, M. C. (2023). A unified approach to demographic data collection for research with young children across diverse cultures. Developmental Psychology. https://doi.org/10.1037/dev0001532

Singh, L., Cristia, A., Karasik, L. B., Rajendra, S. J., & Oakes, L. M. (2023). Diversity and representation in infant research: Barriers and bridges toward a globalized science of infant development. *Infancy*, *28*(4), 708–737. https://doi.org/10.1111/infa.12547

Sperry, D. E., Sperry, L. L., & Miller, P. J. (2019). Reexamining the verbal environments of children from different socioeconomic backgrounds. *Child Development*, *90*(4), 1303–1318. https://doi.org/10.1111/cdev.13195

Tomasello, M., & Mannle, S. (1985). Pragmatics of sibling speech to one-year-olds. *Child Development, 56*(4), 911–917. https://doi.org/10.2307/1130104

Vogt, P., Masson-Carro, I., & de Jong, C. (2021). *Multimodal interactions among infants in three radically different learning environments. Journal of Child Language, 48*(4), 787–808. https://doi.org/10.1017/S0305000921000343

Warstadt, A., Choshen, L., Mueller, A., Williams, A., Wilcox, E., & Zhuang, C. (2023). Call for Papers--The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*.

World Values Survey Association. (2020). World Values Survey: Round Seven - Country-Pooled Datafile Version. Madrid, Spain & Vienna, Austria: JD Systems Institute. Retrieved from https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp.

Zambrana, I. M., Ystrom, E., & Pons, F. (2012). Impact of gender, maternal education, and birth order on the development of language comprehension: A longitudinal study from 18 to 36 months of age. *Journal of Developmental & Behavioral Pediatrics*, *33*(2), 146–155. https://doi.org/10.1097/DBP.0b013e3182423aab