How WEIRD-biased is CHILDES’ data on children’s linguistic input

Camila Scaff*1,2, Georgia Loukatou*2, Alejandrina Cristia2, & Naomi Havron3

1 University of Zurich, Institute of Evolutionary Medicine (IEM), Switzerland

2 PSL University, Laboratoire de Sciences Cognitives et de Psycholinguistique (ENS, EHESS, CNRS, DEC), France

3 Haifa Univerity, Israel

Abstract

In recent years, the importance of estimating demographic biases in research has become apparent. Here we provide a systematic review of the CHILDES archive, the major source of data on naturalistic recordings of children’s linguistic environment. We analyzed the archive at the country and corpus level for four dimensions considered central for language learning: SES, urbanization, family structure and language. We compared these descriptive statistics to world statistics to assess whether the archive was biased in terms of the demographics of the countries represented and the families recorded within them. We found that at the country level, the 47 countries from which there were recordings in CHILDES overrepresented countries with higher educational level; were more urban; and had smaller households with less children. At the corpus level, middle- and higher-class participants were over-represented in relation to the statistics of their own countries. Corpora also included more educated families, with academics being especially 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 the respective countries. In terms of family structure, nuclear families were more prevalent than in the countries the data was collected in, and - surprisingly - children with no siblings appeared to be under-represented. Last, we found that corpora were linguistically diverse, but we estimate that data in CHILDES under-represents bilingual and multilingual households. We conclude that when generalizing from analysis of data obtained from CHILDES, researchers should acknowledge the potential biases of the archive.

*Keywords:* childes, verbal input, infant-directed speech, language, weird

*Word count:* 8,300 words

How WEIRD-biased is CHILDES’ data on children’s linguistic input

Research highlights \* We examined CHILDES, the major source of data on children’s linguistic environment for potential sampling bias in terms of SES, urbanization, family structure and languages. \* We found that the 47 countries present in CHILDES overrepresented rich, educated, urbanized countries with small nuclear families. \* Within these countries, corpora overrepresented rich, educated, urban and nuclear families - and single-child families, as well as bilingual families were underrepresented. \* Interpretation of studies based on CHILDES should acknowledge these biases.

This paper provides an analysis of the CHILDES database, aiming to assess any potential biases in demographic sampling. We examine the database through the lenses of different social and demographic factors (SES, urbanization, family structure, and language) that influence language learning. 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. Naturalistic recordings provide insight into how children acquire language in everyday contexts and capture the richness and complexity of language use in everyday conversations. They also constitute a valuable and ecologically valid source of data necessary for computational models seeking to reverse engineer language acquisition or simulate natural language development. Additionally, they form the foundation for many influential concepts in developmental science. For instance, they helped establish correlations between caregiver behavior, home environment, and child development, such as the link between socio-economic status and language and literacy (Hoff, 2013). Last, they inspire theories that help understand the connection between language input quantity and language development (Christiansen et al., 2022). However, naturalistic recordings may also contain potential biases. One notable concern is that participants might not represent the diverse socio-economic backgrounds, cultures, or linguistic environments children typically experience. As a result, the generalizability of such research may be limited, necessitating cautious interpretation. Researchers in many fields are increasingly aware of the bias towards WEIRD populations (Western, Educated, Industrial, Rich, Democratic, Henrich 2010) in the samples they study (Nielsen et al., 2017; Moriguchi, 2021; see Cychosz & Cristia, 2022, 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 sampled 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 populations and languages studied. Similarly, Blasi et al. (2022) emphasize the potential consequences of generalizing observations derived solely from English speakers and how fit they are to represent our entire human species. [TRANSITION] 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 the development of child language, particularly related to naturalistic recordings: Socioeconomic Status, Urbanization, Family Structure, and Language. ### Socioeconomic Status Decades of research have examined the linguistic differences among families with varying socio-economic status (SES). In the developmental literature, this is primarily indexed by parents’ education (Ensminger, Fothergill, Bornstein, & Bradley, 2003; Hoff, 2003a), but can also be indexed otherwise, such as by parental income, occupation, or a composite measure of those three (e.g. Hollingshead, 1975). 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 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. It should be noted, however, that some of this literature has been found to reflect majority biases, including what kinds of language input are counted, and what features of linguistic experiences are valued (Sperry, et al., 2019; Scaff et al., 2023). Without desiring to take a stance on how SES and language environments relate to each other, we merely indicate here that SES is undoubtedly one of the factors that has been repeatedly studied in the context of early language acquisition, including in terms of input description. For example, Hoff (2003a) compared the speech of high- versus low-SES American mothers. College-educated parents produced more utterances to their child, with more diverse vocabulary, and longer phrases, and 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; for meta-analyses supporting the link; and Bergelson et al., 2023 for a large-scale study finding non-significant SES effects).

### Urbanization.

Urbanization is the process of moving from rural to urban areas along with noticeable changes in job opportunities and living conditions. It involves the growth and development of cities, leading to increased access to infrastructure and amenities. Studies indicate that children exposed to higher levels of noise pollution in urban areas exhibit a reduction in cortical thickness in the left IFG, an area of the brain associated with language development (Simon et al., 2022). Within the general theoretical framework of language socialization, there have been proposals suggesting that societies varying in their urbanization process have differing views and values about the role of children in conversations, and more generally in the community (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. These three groups differ in terms of their goals for children, with urban families aiming for child psychological independence, rural families for child physical autonomy and interdependence, and hybrid families aiming for some mix across these values. Vogt et al. (2020) employ this conceptual classification to interpret their results on multimodal language use across three samples, namely urban Dutch, urban Mozambique, and rural Mozambique, finding 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.

### Family structure.

Family structure refers to the arrangement within a household, forming the basis of a family unit. This dimension encompasses aspects such as the number of siblings, birth order, the number of caregivers in the household, and the number of individuals sharing or competing for household resources (including caregiving attention); each of which has a significant 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 example, birth order effects reveal that children with older siblings show lower language skills than first-born children in various cultures (e.g., Pyere et al., 2016 in France; Havron et al., 2022 for Singapore; Zambrana, et al., 2012 for Norway). [Another example], in middle-class Euro-American families, parents typically assume primary responsibility for children, often focusing on 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 Turkish families described by Isleyen (2021), may adopt a different approach, with nuclear families living in separate apartments but sharing common spaces and caregiving responsibilities, resulting in extensive support networks. [Importance of family size. BUT Family size does not seem to affect language acquisition when controlling for factors such as SES and birth order (see Blake, 1981). ### 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, as variations in language exposure and language use among different groups allow to explore how purely linguistic factors shape and influence the development of language skills in children. Another major dimension and entire sub-field in developmental science is the study of bilingualism or multilingualism (for an introductory review see XX, also McCabe et al., 2013). We refer to lingual status as the variable that distinguishes monolingualism from bilingualism and other cases.

### The current study.

Here, we provide a systematic analysis of the naturalistic speech corpora of the CHILDES database by quantifying the diversity of each dimension presented above (see Tables 1 and 2). Though some of these dimensions overlap with each other (See Figures SM.X and SM.x in the supplementary materials for illustrations), we decided to illustrate each, as best as possible, independently. Although valuable insights can be gleaned from corpora on clinical populations and elicitation tasks, we deliberately excluded them from our review. This is because they introduce additional biases. For example, the inclusion of clinical populations may highlight differences in the prevalence and access to diagnosis of language disorders in different regions or countries. Additionally, the structure of elicitation tasks may not accurately reflect spontaneous language use in everyday conversations. Our systematic analysis follows three steps. First, we screened the CHILDES database, excluding the clinical and task-driven recordings. Second, we extracted information related to the four central factors in language acquisition described in the Introduction. Finally, we follow Ghai’s (2022) recommendation to improve the description of diversity in behavioral sciences by looking at different levels: from a macro-level, with broad country comparisons to a corpus-level, where we delved into individual corpora to gain a more detailed and nuanced understanding of the data.

# Methods

Analyses and visualizations were carried out using R (version 4.1.2, R Core Team, 2020) and ggplot2 (Wickham, 2016). Data, scripts, and online Supplementary Materials are available on OSF [<https://osf.io/q9w82/?view_only=a013f1b25b8c4556b8248f12870402c9>].

### Inclusion criteria.

We excluded the following corpora: a) clinical populations or non-typically developing children, b) structured tasks such as toy narratives, personal narratives, frog stories, picture or movie descriptions, structured storytelling, and other elicitation tasks; c) only child or adult speech without a conversational partner; and d) non-naturalistic setups, such as recordings conducted exclusively in a lab environment or in a diary format. ### Screening Following a thorough examination of each corpus, which involved reviewing the corpus description available on the CHILDES website, checking for any accompanying references such as articles, book chapters, or dissertations, and conducting spot-checks on associated transcripts, we identified 180 corpora that met our inclusion criteria mentioned above. For a detailed breakdown of the included corpora, please refer to the flowchart in Supplementary Materials.

### Macro-level information.

Table 1:

\*\*

| Dimension | Macro.level.variable | Source |
| --- | --- | --- |
| SES | Percent of the population completing lower secondary school\* | Data from 2007-2015; Our World in Data, 2022a |
| SES | GDP per capita (log 10) | World Development Indicators (WDI) from the World Bank, available via the WDI package in R (Vincent Arel-Bundock, 2021) |
| Urbanization | Percent of the population living in urban areas | WDI from the World Bank. It represents the proportion of the population that was rural, so we estimated the complement to 100% to conceptually align this variable to the others. We employed indicators for the year 2011 because this was the year in which we had the maximum amount of data available for those variables. |
| Family structure | Average household size | UN database on Household Size and Composition 2022 (data from 2000-2022 United Nations, 2022) |
| Family structure | Average number of member under 15 in households | UN database on Household Size and Composition 2022 (data from 2000-2022 United Nations, 2022) |
| Language | NA | NA |

The goal of the country-level analysis is to assess the representativeness of our sub-sample of CHILDES against world statistics.. The following variables were used for this level of analysis, all derived from official sources including the World Bank, Our World in Data (WDI), and the United Nations (UN). Add sentence about why we dont have something on multilingualism at this level: not good data at the country level for this as some languages or dialects are not be counted blabla…

### Corpus-level information.

Table 2:

*Definition of the corpus-level variables*

| Dimension | Corpus.level.variable | Definition.or.levels | N.of.corpus…missing. |
| --- | --- | --- | --- |
| SES | Parents’ socio-economic status | Low, mid and/or high |  |
| SES | Parents’ education level | Highest level of education completed: Primary, High school, Undergraduate, Masters, PhD |  |
| SES | Parents’ occupation | Parents’ activity or profession |  |
| Urb | Type of community | Urban, rural, mixed |  |
| Family structure | Household composition | Whether the family was composed primarily of caregivers and children (nuclear) or it also included other family members: grandparents, cousins, uncles, and aunts (extended) |  |
| Family structure | Percent children with sibling(s) | The percentage of children in the corpus who had at least one sibling |  |
| Family structure | Average number of siblings | How many siblings children had on average (including children with zero siblings) |  |
| Language | Language(s) spoken | Which languages were spoken in the transcripts and by the family |  |
| 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) |  |

The goal of the corpus-level analysis was to assess the representativeness of our sub-sample of CHILDES in a finer-grained way across our four dimensions: SES, urbanization, family structure and languages (See Table 2 for the complete list of variables and definitions). Data was extracted from the provided sources mentioned in CHILDES such as articles, book chapters, dissertations, and transcripts to pre-fill the categories. Corpus curators were contacted to request missing or incomplete information, and an online table was provided to facilitate data entry. Over a third of the contacted curators (XX corpora, XX%) provied additional data or confirmed missing information. Unfortunately, curators for 50 corpora (XX%) could not be contacted due to unresponsive email addresses (38) , or the curator’s passing (12).

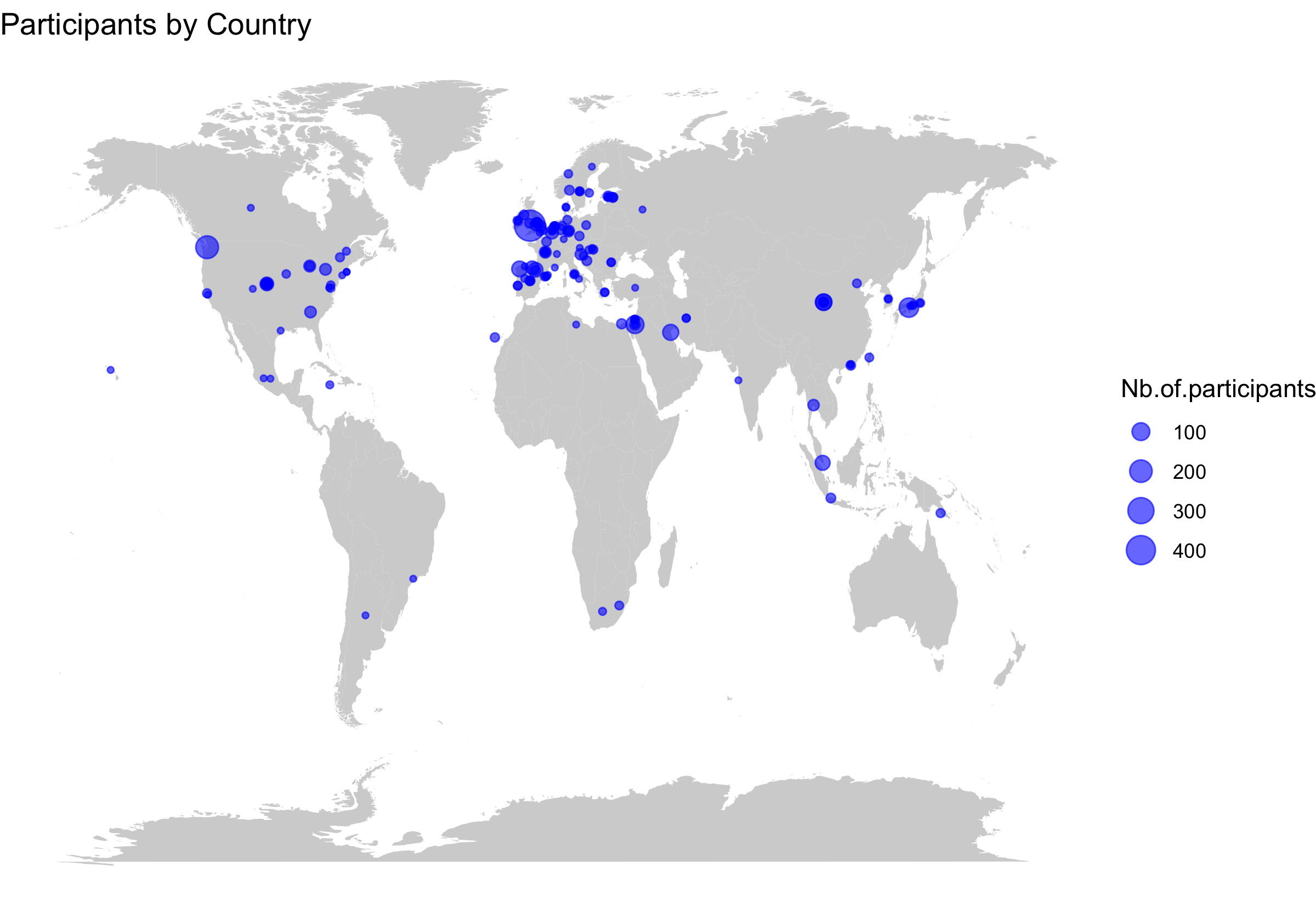
# Results

The 180 corpora included in this study represent 47 different countries across all populated continents. Out of the 48 different countries or areas included in the sample, 28 of them belong to the OECD (Organisation for Economic Co-operation and Development). This corresponds to 151 out of 180 corpora or to 84% of this sub-sample of CHILDES. OECD countries represent 19.5% of world countries. See Table 3 and Figure 1 for distribution of corpora in the globe.

Table 3:

*Descriptives*

| Continent | N.corpora | N.total.children | N.children.growing.in.continent |
| --- | --- | --- | --- |
| Africa | 3.00 | 16.00 | X |
| Asia | 32.00 | 602.00 | X |
| North America | 34.00 | 177.00 | X |
| Latin America | 5.00 | 6.00 | X |
| Europe | 105.00 | 1,091.00 | X |
| Oceania | 1.00 | 5.00 | X |



*Figure* *1.*  World map showing corpus. Size proportional to number of children per corpus

We assessed the extent to which the countries with data in our sub-sample of CHILDES were a representative sample of countries in the world. Density plots are portrayed in Figure 2.

By comparing our sub-sample of CHILDES to the world statistics using unpaired samples t-tests without assuming equality of variance (Welch’s t). Countries in our sub-sample of CHILDES had a higher proportion of the population completing lower secondary school than the world wide sample (% compl. LSS, t(101.59) = 5.79, p < .001); they were more urban (% urban, t(79.77) = 3.51, p < .001); richer (log GDP per capita, t(103.29) = 6, p < .001) and had smaller households (average household size, t(118.8) = 7.116, p p < .001) and less average number of members under the age of 15 (average under 15 size, t(26.118) = 5.01, p p < .001).

## Corpus-level

We then investigated at the corpus level the variables described in Table 2. The information collected allows us to have more detail about the characteristics of the individual families that compose our sub-sample of CHILDES. However, it is important to acknowledge that for the majority of the corpora, this information was missing, not known, or not provided.

### SES.

Of the 102 corpora, 5 were described as having low SES; 16 were described as spanning both lower and middle or higher SES; and 81 were described as middle or higher SES exclusively (80%). Given that most countries represented in our sub-sample of CHILDES are in the OECD, we can compare this proportion with the proportion of the population in these countries that are middle class. According to a 2016 report, “Almost two-thirds of people live in middle-income households in OECD countries”, for whom “household net income [is] between 0.75 and 2 times the median”. Thus, middle and higher-class participants appear to be over-represented in our sub-sample of CHILDES data.

#### Education.

Four corpora were described as “diverse”, without clarifying the range of education covered. Of the remaining 76 corpora, 3 had at least some parents with primary-level education; 9 had parents with secondary school education as the lower bound of the education range, and a further 6 had some college as the lower bound. Thus, 72.5% of the corpora (N = 58) include children whose parents had at least a graduate, if not a postgraduate, degree. The collected corpora do not accurately represent the demographics of the countries they were obtained from. For instance, while the U.S. Census Bureau reported that only 36% of the adult population held a bachelor’s degree or higher in 2020, our data indicates that 100% of the corpora from people in the United States had a college education or higher. As seen in Figure XX, the same result is seen for corpus from China, Japan, Israel, South Korea, Italy, Hungary, Mexico, The Netherlands, Poland, and Switzerland, … where 100% of parents with data are college educated or above. Thus, it seems that our sub-samples of CHILDES are very skewed toward higher-educated parents.

#### Occupation.

Professions were overall varied, but 50% of the corpora contained parents who were described as one of the following: Masters or Ph.D students, professors, linguists, researchers, scientists, or academics. Some of the other professions reported were ….

### Urbanization.

58 corpora were described as industrialized or urban, and an additional one as both rural and urban. Only 7 corpora were described as farming or rural. In these same countries, the proportion of the population residing in urban settings was 76%, suggesting that samples were not representative of their countries in terms of rural versus urban settings either.

### Family structure.

#### Nuclear households.

As for household composition, 60 corpora (88%) were based on nuclear families; 6 extended families(XX%); and in 2 corpora the structure was varied (XX%). We compare side by side the average per country of the nuclear households (the sum of the percentages of couples with children households, and single parents with children households; United Nations, 2022).

#### Sibling presence.

A majority of corpora in our sub-sample of CHILDES include children who have siblings. In fact, only 28% of corpora (among the 93 corpora having information on siblings) were constituted exclusively of children with no siblings, and the remaining had at least one sibling, with the overall average being 0.8 siblings. Since 83% of countries in CHILDES are in the OECD, we draw a comparison point for such countries: 46% of children had no siblings in OECD countries according to 2015 data. In this sense, children with no siblings appear to be under-represented in CHILDES.

### Languages.

We had two variables of interest here: language spoken in the corpus and lingual status. A total of 62 different languages or language combinations (for bilingual and multilingual children) were reportedly spoken in the corpora.

Table 4:

*Total number of participants per language,*

| Monolingual\_Language | Total\_Participants | Multilingual\_Languages | Total\_Participants\_Multilingual |
| --- | --- | --- | --- |
| Afrikaans | 2.00 | Dutch/English, Dutch/French | 35.00 |
| Arabic (Egyptian or Kuwaiti) | 80.00 | Dutch/Italian | 4.00 |
| Basque | 46.00 | Spanish/Catalan | 6.00 |
| Cantonese | 8.00 | Spanish/Galician | 66.00 |
| Catalan | 17.00 | English/Cantonese | 9.00 |
| Cree | 1.00 | English/French | 12.00 |
| Croatian | 3.00 | English/Hebrew | 1.00 |
| Czech | 6.00 | English/Japanese | 1.00 |
| Danish | 2.00 | English/Japanese/Danish | 1.00 |
| Dutch | 23.00 | English/Mandarin | 55.00 |
| English | 210.00 | English/Mandarin/Cantonese | 11.00 |
| Estonian | 32.00 | English/Russian | 22.00 |
| Farsi | 5.00 | English/Spanish | 12.00 |
| French | 46.00 | French/Russian | 1.00 |
| German | 46.00 | Hungarian/Catalan/Spanish | 1.00 |
| Greek | 6.00 | Hungarian/Farsi/English | 2.00 |
| Hebrew | 122.00 | German/Spanish | 9.00 |
| Hungarian | 8.00 | Italian/Japanese | 1.00 |
| Icelandic | 1.00 | Italian/German | 2.00 |
| Indonesian | 8.00 | Portuguese/Swedish/English | 3.00 |
| Irish | 7.00 | NA | NA |
| Italian | 8.00 | NA | NA |
| Jamaican | 2.00 | NA | NA |
| Japanese | 148.00 | NA | NA |
| Korean | 4.00 | NA | NA |
| Mandarin | 157.00 | NA | NA |
| Norwegian | 11.00 | NA | NA |
| Nungon | 5.00 | NA | NA |
| Polish | 4.00 | NA | NA |
| Portuguese (Brazilian or European) | 9.00 | NA | NA |
| Romanian | 6.00 | NA | NA |
| Russian | 2.00 | NA | NA |
| Serbian | 8.00 | NA | NA |
| Sesotho | 4.00 | NA | NA |
| Slovenian | 20.00 | NA | NA |
| Spanish | 75.00 | NA | NA |
| Swedish | 9.00 | NA | NA |
| Taiwanese | 4.00 | NA | NA |
| Tamil | 1.00 | NA | NA |
| Thai | 18.00 | NA | NA |
| Turkish | 1.00 | NA | NA |
| Welsh | 475.00 | NA | NA |

About a third (32%) of the included corpora that had available data for this variable (N = 110) were not monolingual. It is hard to find reliable estimates of the percentage of the population which is not monolingual in the world or in the countries represented in our sub-sample of CHILDES, but for instance, in Europe in 2016, 65% of adults reported knowing multiple languages (Eurostat, 2022). According to such estimates, even if samples are linguistically diverse, it would appear that input data in our sub-sample of CHILDES under-represents bilingual and multilingual households.

# Discussion

The current paper provides a systematic review of naturalistic recordings from the CHILDES database. 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 influence early language acquisition. We first analyzed the database by the country where the corpus was recorded (country-level analysis), and then using individual corpus information (corpus-level analysis). We aimed to address two descriptive research questions: (1) According to the country-level variables, to what extent do countries in this sub-sample of CHILDES represent countries in the world? and (2) According to corpus-level variables, to what extent do corpora in this sample of CHILDES represent their own countries? The systematic review included 180 corpora from 47 different countries, mainly representing OECD countries (84%). OECD countries have higher income on average than non-OECD countries (CITE). The analyses at both country and corpus levels revealed several measures where the corpora do not represent the world population across SES, urbanization, family structure, and linguistic dimensions within the CHILDES database. Despite a remarkable diversity in terms of languages and countries, naturalistic recordings seem to encompass quite similar socio-ecological backgrounds. The socioeconomic status (SES) dimension encompassing education, income, and occupation, shows a stark bias of our CHILDES sub-sample from global statistics. Specifically, our represented countries demonstrate a higher percentage of individuals completing lower secondary school compared to global averages. A more detailed examination reveals 78% of the dataset represent families where parents attained at least a college education. Data for XX out of 49 countries exclusively reported parents with college-level education or above, whereas for some countries only some primary or secondary school parents are represented. Furthermore, these countries exhibit higher Gross Domestic Product (GDP) per capita, paralleling our findings on SES at the corpus level, where 80% of the reported SES are middle or high SES. Taken together, it is clear that naturalistic input samples in CHILDES represent higher SES and educated countries and households. 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.

When examining the urbanization dimension, we see that in the countries of this sub sample of CHILDES the majority of the population resides in urban areas. And that X% of corpora were urban, as opposed to XX in the world population, and only X corpora were exclusively rural. It has been shown that urban and rural families often have different expectations about child development (REF),and that urban settings have more access to services such as daycares, or playgrounds, creating particular early language environments that research has shown that affects XYZ in language development. On the other hand, recent research also shows the negative effects of urban areas for health and particularly, related to noise pollution, for which links with early language development have been found - in this respect too, the biased data could be problematic. For the family-structure dimension, CHILDES country level analysis shows that the represented countries had smaller households and on average, and fewer family members under the age of 15 when compared to the world countries. These results also manifest at the corpus level, showing an overwhelming number of nuclear families as opposed to extended families. This limits our understanding of how diverse caregiving structures, such as alloparental involvement, affect language acquisition. While multi-child families are well represented in CHILDES, the absence of siblings’ speech in some transcripts (e.g., Loukatou et al., 2022) could hinder insights into typical daily language interactions between them. Furthermore, the dataset’s lack of birth order information restricts the analysis of sibling dynamics’ influence on language acquisition. This scenario underscores the need for broader exploration of non-parental caregiving roles and sibling relationships, typically overlooked in studies focusing on predominantly WEIRD households. Finally in terms of language, in naturalistic CHILDES corpora, the most represented language family is XX. Additionally, for XX out of 180 corpora, English is the recorded language. Thus, English does not appear over-represented in terms of number of corpora (but see Christiansen et al., 2022, for evidence that it may be in terms of overall quantity of data). In terms of lingual status, the majority of the data comes from monolingual children, which is not representative of the world’s population (CITE).

We chose to focus on four central factors in language acquisition, but others may well be relevant. For example, Hofstede (2001) conceptualized dimensions such as individualism, masculinity, and indulgence. Moreover, given that most samples are nuclear families with well-educated parents living in urban sites, it will be impossible to tease apart these three dimensions, which are in effect confounded in these CHILDES corpora (and probably in the world, see Supplementary Materials SMX). Even in a case where we do have some variation, namely the presence of siblings, it would be hard to understand the intersectionality of such effects in naturalistic input samples. 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 are not always available to them or were not considered relevant before. However people re-using the data should try at least to make an effort to inform themselves about the characteristics of the people recorded as they could be of interest to interpret findings. Researchers should acknowledge the database’s limitations in representing diverse naturalistic language environments, especially when exploring socioeconomic, urbanization, or family structure influences on language acquisition. In addition, as AI technologies and other types of scraping tools become prevalent, it is important to acknowledge the socio-ecological landscape of this repository in order to keep in mind what are the generalizations that automatic tools might arrive at, and how they represent the overarching mechanisms of language acquisition. 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 could not be viewed as representative of their own countries. We also noted low variability along certain dimensions (saliently family structure) and general confounds across dimensions. Our discussion led us to argue for the systematic inclusion of certain descriptors in speaker-level and corpus-level metadata, in order to track and quantify both diversity and representativity (see Sup Mat XX).

# References