Supplementary Materials to Demographic Biases in Naturalistic Language Recordings in the CHILDES Database

## SM1: Sources used for Figure 2

The purpose of this figure is to illustrate how representative the countries in our CHILDES sample are compared to global distributions. We assessed socioeconomic status (SES) using two key indicators: education and income. For education, we used data from Our World in Data (2022), which compiles information from various official sources. Specifically, we obtained data on the proportion of the population that had completed lower secondary education (from 2007-2015). For income, we utilized the World Bank's World Development Indicators (WDI) via the WDI package in R (Arel-Bundock, 2021), focusing on Gross Domestic Product (GDP) per capita (log-transformed to base 10) in 2017 US dollars adjusted for purchasing power parity, reflecting the relative wealth of countries.

To measure urbanization, we used the percentage of the population living in urban areas, sourced from the World Bank's WDI. Since this indicator initially represented the rural population proportion, we adjusted it to represent urbanization by calculating the complement to 100%. We used data from 2022, as it offered the most comprehensive information for this variable.

For family structure, we referred to the average household size, based on data from the UN’s Household Size and Composition database (2000-2022; United Nations, 2022). This measure reflects the average number of individuals per household, providing insight into family structure across different countries.

## SM2: Higher SES families are overrepresented in our sample of CHILDES given source countries

As mentioned in the main text, the vast majority of families were of middle or higher SES. Since 84% of corpora were collected in OECD countries, we use OECD data for comparison. While less than 66% of families in OECD countries are middle-to-high SES, over 80% of our sample falls into this category. Moreover, OECD countries generally have higher SES populations compared to the global average, suggesting that our CHILDES sample likely overrepresents middle-to-high SES families even more on a global scale.

## SM3: Caregivers with high levels of education are overrepresented in our sample of CHILDES given source countries

To check whether the levels of education were reasonable, we calculated the proportion of corpora in each country for which all caregivers had completed college, setting aside the corpus collected in two countries, for a total of 27 countries. This percentage varied between 0 (Greece, Lesotho, Serbia, or Portugal) and 100% (for 16 countries). In contrast, for the same 27 countries, the proportion of the population aged 25-65 years with completed tertiary education did not go above 30% (data from 2010). Zooming into one particular case, 100% of families in the corpora collected in the United States had a college education or higher, which is true of only 36% of US adults (U.S. Census Bureau, 2020).

In fact, 50% of the samples contained parents who were described as master’s or Ph.D. students, professors, linguists, researchers, scientists, and academics (if we additionally count psychologists and speech therapists, this represents 62% of the corpora). To give an idea of the extent to which this is not representative, consider the fact that in 2020, 6% of the American population would be included in that list of professions. Similar data is hard to find for all countries represented in CHILDES, but we suspect that the proportion of scientists, professors, Masters, and PhD students found in most other countries will be the same or lower.

## SM4: Urban families are overrepresented in our sample of CHILDES given source countries

## To check whether the levels of urban sampling were reasonable, we calculated the proportion of corpora in each country for which all families were urban, setting aside the corpus collected in two countries, for a total of 26 countries. This percentage varied between (Jamaica, Lesotho, Papua New Guinea) and 100% (for 20 countries). For the same 26 countries, the proportion of the population that was urban varied more continuously, ranging from 14 (Papua New Guinea) to 100% (Kuwait). Among the 20 countries for which 100% of families were urban, this percentage varied between 43 and 100%. For the remaining 6 countries, it went from 14 to 83%.

## SM5: Singletons may be underrepresented in our sample of CHILDES given source countries

#### Since 84% of the countries in CHILDES are in the OECD, we draw a comparison point for such countries: among households with children in the OECD, roughly 46% of children were singletons (OECD, 2015). In the sub-sample of OECD CHILDES corpora with information about siblings, we see that 29% (n = 23) of corpora focus exclusively on singletons. In this sense, singletons are underrepresented in CHILDES. However, we are not certain of this interpretation, because of the way data were coded: It is not the case that the remaining 71% of corpora contain only children with siblings, but rather that at least some of the children in the other corpora had siblings. Therefore, to be truly certain, one would have to do an analysis based on individual included children, which we were not able to do at present because such information was not always systematically reported for all children in each corpus.

## SM6: Languages represented

Monolingual corpora contained data in: Afrikaans, Arabic (Egyptian or Kuwaiti), Basque, Cantonese, Catalan, Cree, Croatian, Czech, Danish, Dutch, English, Estonian, Farsi, French, German, Greek, Hebrew, Hungarian, Icelandic, Indonesian, Irish, Italian, Jamaican, Japanese, Korean, Mandarin, Norwegian, Nungon, Polish, Portuguese (Brazilian or European), Romanian, Russian, Serbian, Sesotho, Slovenian, Spanish, Swedish, Taiwanese, Tamil, Thai, Turkish, and Welsh.

The non-monolingual corpora contained data in: Dutch/English, Dutch/French, Dutch/Italian, English/Cantonese, English/Dutch, English/French, English/Hebrew, English/Japanese, English/Japanese/Danish, English/Mandarin, English/Mandarin/Cantonese, English/Russian, English/Spanish, French/Russian, German/Spanish, Hungarian/Catalan/Spanish, Hungarian/Farsi/English, Italian/German, Italian/Japanese, Portuguese/Swedish/English, Spanish/Catalan, and Spanish/Galician.

## SM7: Monolinguals are over-represented given estimates of non-monolingualism in the world

It is hard to find reliable estimates of the percentage of the population that grows up in bilingual or multilingual homes in individual countries. This is partly because estimates rely on adult data which include second language learners, not exposed to an additional language in their home environment. For instance, in Europe in 2016, 65% of adults reported knowing multiple languages (Eurostat, 2022), but most of them probably did not grow up in bilingual homes. Still, some estimates claim that around half or two-thirds of children are raised bilingual or multilingual (Grosjean, 2024).

## SM8: Acknowledgments

We would like to thank the corpora curators: Airi Kapanen, Alan Cruttenden, Alison Henry, Aliyah Morgenstern, Amy Strekas, Amye Warren-Leubecker, Ana Isabel Ojea Lopez, Ana Lúcia Santos, Ana Maria Guimarães, Andrea Biró, Andrea Feldman, Angela Grimm, Ann Peters, Anna Chromá, Anna Theakston, Anne Van Kleeck, Anne-Marie Schaerlaekens, Annick De Houwer, Antje van Oosten, Aparna Nadig, Astrid Klammler, Aurora Bel Gaya, Aviya Hacohen, Ayhan Aksu Koç, Barbara Davis, Barbara Pearson, Bernadette Plunkett, Bernd Möbius, Bob Jones, Bob Wilson, Brian MacWhinney, Britta Lintfert, Carina Koroschetz, Carmen Silva-Corvalán, Caroline Rowland, Catherine Snow, Cécile De Cat, Charles Watkins, Chiara Roggero, Chien-ju Chang, Christian Champaud, Christiane von Stutterheim, Christina Gildersleeve-Neumann, Christine Howe, Claartje Levelt, Claudine Hammelrath, Colleen Huebner Morisset, Conxita Lleo, Conxita Lleó, Cornelia Hamann, Darinka Anđelković, David Dickinson, David Gil, Dmitar Popov, Donella Antelmi, Donna Jackson-Maldonado, Donna Thal, Dorit Ravid, Eithne Guilfoyle, Ekaterina Protassova, Elena Lieven, Elisabet Serrat Sellabona, Eliseo Diez-Itza, Elizabeth Bates, Elizabeth Nixon, Eon-Suk Ko, Eva Bar-Shalom, Eve Clark, Evelien Krikhaar, Feyza Altinkamis, Filip Smolik, Folkert Kuiken, Francisco De Lacerda, Frank Wijnen, Fred Genesee, Frenette Southwood, Gaja Jarosz, Gerard Bol, Gerardo Aguado Alonso, Ghada Khattab, Gina Conti-Ramsden, Gisela Szagun, Giuseppe Cappelli, Gordana Hržica, Gordon Wells, Habibeh Samadi, Hanna Batoréo, Hannah Sarvasy, Harriet Jisa, Heather Goad, Heba Salama, Heidi Feldman, Heike Behrens, Helen Körgesaar, Hervé Hunkeler, Hintat Cheung, Hiro Yuki Nisisawa, Hrafnhildur Ragnarsdóttir, Hye-Ree Ghim, Igor Žagar, Iliana Reyes, Inge Zink, Ioana Goga, Isabelle Barrière, Isabelle Maillochon, Jacqueline Sachs, Jacqueline van Kampen, Jane Herbert, Jane S. Tsay, Janet Bang, Jan Edwards, Javier Aguado Orea, Jean Berko Gleason, Jean Quigley, Jean-Adolphe Rondal, Jeannine Goh, Jeroen Aarssen, Jing Zhou, Jody Tommerdahl, Joe Pater, Johanna Nicholas, Jóhanna Thelma Einarsdóttir, Johanne Paradis, John Neil Bohannon III, Jordan Zlatev, José L. Linaza, Ju-Yeon Ryu, Judit Navracsics, Julian Pine, Julie Brittain, Julie McMillan, Jürgen Weissenborn, Juana Liceras, Kaja Kohler, Karina Hess Zimmermann, Karme Beek, Katerina Palasis, Katherine Demuth, Katherine Nelson, Kathy Post, Keith Sawyer, Kim Plunkett, Klaus Wagner, Klára Matiasovitsová, L. Haggerty, Laetitia de Almeida, Larisa Avram, Larry F. Guthrie, Leonor Scliar-Cabral, Liliana Tolchinsky, Linda Kelly, Linhui Li, Lise Menn, Livia Tonelli, Lois Bloom, Lori Van Houten, Lorraine McCune, Lynn S. Bliss, Madalena Cruz-Ferreira, Madeleine Leveillé, Magda Krupa-Kwiatkowska, Magdalena Smoczynska, Maigi Vija, Maja Roch, Manuela Wagner, Marc Bornstein, Margaret Deuchar, Maria del Carmen Aguirre Martínez, Maria Emma Ticio, María Jesús Pérez-Bazán, Maria João Freitas, Maria-Llanos Luque Sánchez, Marie-Thérèse Le Normand, Mariko Hayashi, Marilyn Vihman, Marta Fernández Vázquez, Martha Shiro, Marty Demetras, Mary Ann Evans, Mary Beckman, Mary Erbaugh, Masayuki Yokoyama, Mats Andrén, Max Miller, Megan Devlin, Melanie Soderstrom, Melissa Redford, Melita Kovacevic, Michael Brent, Michael Forrester, Michelle McGillion, Michelle White, Milagros Fernández Pérez, Miquel Serra, Mirco Fasolo, Mireas Llinas, Mireia Llinàs-Grau, Mitsuhiko Ota, Mohamed Lahrouchi, Monique Vion, Myron Korman, Nan Bernstein Ratner, Naomi Hamasaki, Naomi Yamaguchi, Natalia Gagarina, Neil Smith, Neiloufar Family, Nina Gram Garmann, Níria Esteve-Gibert, Norio Naka, Núria Esteve-Gibert, Oksana Bailleul, Ondene Van Dulm, Oralia Rodríguez Arredondo, Outi Bat-El, Pamela Rollins, Patrick Suppes, Paul Fletcher, Paula Fikkert, Péter Bodor, Petra Bos, Petra Hendriks, Petra Sleeman, Pilar Prieto, Rangaswamy Narasimhan, Raquel Fernández Fuertes, Rebecca Burns, Reili Argus, Richard Sprott, Richard Weist, Roberto Soto Valle, Roger Brown, Ron Gillam, Rosa Graciela Montes, Roy Higginson, Ruth Berman, Sam Leung, Sanne Kuijper, Seba Al-Hindawy, Shaima AlQattan, Sharon Inkelas, Silvia Nieva, Silvia Romero Contreras, Sirli Zupping, Sophie Kern, Sotaro Kita, Stan Kuczaj, Stephen Matthews, Steven Gillis, Sudaporn Luksaneeyanawin, Susan Ellis Weismer, Susan Gelman, Susan R. Braunwald, Susana Correia, Susana López Ornat, Susanne Miyata, Sven Strömqvist, Takeo Ishii, Tamirand De Lisser, Tania Ionin, Teresa da Costa, Thea Cameron-Faulkner, Thomas Doukas, Thomas Lee, Tina Ringstad, Twila Tardif, Ursula Stephany, Valentin Remedi, Victoria Marrero, Virginia C. Gathercole, Virginia Gathercole, Virginia Valian, Virginia Yip, William Hall, William Snyder, Xiangjun Deng, Yasuhiro Shirai, Yonata Levy, Yoshiki Ogawa, Yow Wei Quin, Yvan Rose, Zhang Yibin.

## References

Arel-Bundock, V. (2021). WDI: World development indicators and other World Bank data. R package version, 2(4).

Eurostat (2022). Foreign language skills statistics. Available from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Foreign\_language\_skills\_statistics, last visited August 17, 2024.

Grosjean, F. (2024). *Bilingual and Multilingual Children: New Findings and Directions.* New York, NY: Oxford University Press.

Our World in Data (2022). Share of the population with at least some secondary education. Available from <https://ourworldindata.org/grapher/share-of-the-population-with-at-least-some-secondary-education>, last visited August 17, 2024. Data downloaded 2022-02-17.

United Nations (2022). Household Size and Composition database. Available from <https://www.un.org/development/desa/pd/data/household-size-and-composition>, last visited August 17, 2024.

U.S. Census Bureau. (2020). Educational Attainment in the United States: 2019. U.S. Department of Commerce. Retrieved from https://www.census.gov.

## Package and environment version

## R version 4.1.0 (2021-05-18)  
## Platform: x86\_64-apple-darwin17.0 (64-bit)  
## Running under: macOS Big Sur 10.16  
##   
## Matrix products: default  
## BLAS: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.dylib  
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib  
##   
## locale:  
## [1] en\_US.UTF-8/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8  
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] RColorBrewer\_1.1-3 rjson\_0.2.21 ggthemes\_4.2.4 GGally\_2.1.2   
## [5] plotly\_4.10.0 ggpubr\_0.4.0 kableExtra\_1.3.4 stringr\_1.5.1   
## [9] scales\_1.2.0 purrr\_1.0.2 tidyr\_1.3.1 dplyr\_1.1.4   
## [13] ggplot2\_3.3.6 papaja\_0.1.2.9000 tinylabels\_0.2.4   
##   
## loaded via a namespace (and not attached):  
## [1] httr\_1.4.3 splines\_4.1.0 jsonlite\_1.8.8 viridisLite\_0.4.0  
## [5] carData\_3.0-5 datawizard\_0.9.1 highr\_0.11 yaml\_2.3.9   
## [9] bayestestR\_0.13.1 pillar\_1.9.0 backports\_1.5.0 lattice\_0.20-44   
## [13] glue\_1.7.0 digest\_0.6.36 ggsignif\_0.6.1 rvest\_1.0.2   
## [17] colorspace\_2.0-3 Matrix\_1.3-3 htmltools\_0.5.8.1 plyr\_1.8.7   
## [21] pkgconfig\_2.0.3 broom\_1.0.6.9000 xtable\_1.8-4 mvtnorm\_1.1-3   
## [25] webshot\_0.5.3 svglite\_2.1.0 emmeans\_1.7.5 tibble\_3.2.1   
## [29] mgcv\_1.8-35 generics\_0.1.3 farver\_2.1.1 car\_3.1-0   
## [33] withr\_3.0.0 lazyeval\_0.2.2 cli\_3.6.3 magrittr\_2.0.3   
## [37] effectsize\_0.8.6 estimability\_1.3 evaluate\_0.24.0 fansi\_1.0.6   
## [41] nlme\_3.1-152 rstatix\_0.7.0 xml2\_1.3.3 tools\_4.1.0   
## [45] data.table\_1.14.2 lifecycle\_1.0.4 munsell\_0.5.0 compiler\_4.1.0   
## [49] systemfonts\_1.0.4 rlang\_1.1.4 grid\_4.1.0 parameters\_0.21.3  
## [53] rstudioapi\_0.13 htmlwidgets\_1.5.4 labeling\_0.4.2 rmarkdown\_2.27   
## [57] gtable\_0.3.0 abind\_1.4-5 reshape\_0.8.9 R6\_2.5.1   
## [61] knitr\_1.48 fastmap\_1.2.0 utf8\_1.2.4 insight\_0.19.7   
## [65] stringi\_1.8.4 Rcpp\_1.0.9 vctrs\_0.6.5 tidyselect\_1.2.1   
## [69] xfun\_0.46 coda\_0.19-4