

How WEIRD-biased is CHILDES data on childrens linguistic input? Supplementary Materials 1

Contents

Supplementary recommendations for CHILDES	1
Further acknowledgments	1
Additional analyses	3

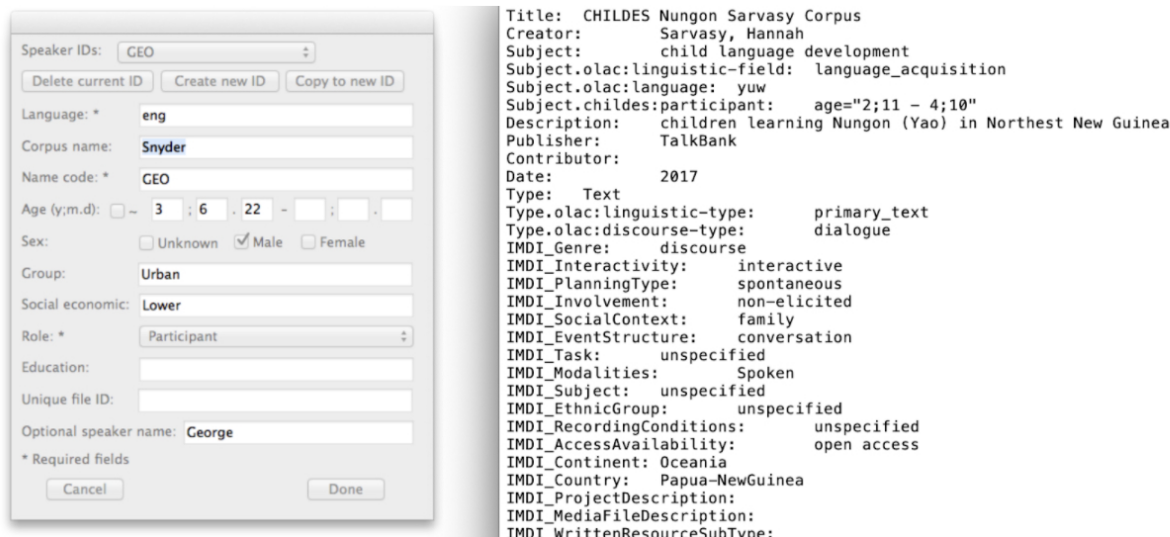
Supplementary recommendations for CHILDES

As represented in Figure 1, CHILDES corpora can have two levels of metadata, one at the speaker level and the other at the level of the corpus as a whole. At present, nearly none of the fields are mandatory and seldom filled in. The standardized (yet optional) fields in speaker-level metadata include several fields that would be relevant to the study of diversity in CHILDES, including: language, age, sex, socioeconomic status (WC for working class, UC for upper class, MC for middle class, LI for limited income), role (relationship to the child), and education (in years). At the corpus level, there are already variables to encode the type of activity recorded (interactivity, planning type, involvement, social context, event structure, and task) as well as information about location (at the continent and country levels). There is, at present, no way to signal the geographic origin of CHILDES data more finely, but if this information were added, then we could assess the extent to which localities where education is widespread are over-represented; and the extent to which unusually well-educated parents are over-represented. Perhaps a reasonable goal is to attempt to match the world’s distribution in terms of simple educational metrics, like having completed lower secondary school.

Incorporating these other metrics would require not only adding fields to the CHILDES’ metadata system, but also developing training materials to ensure that all CHILDES contributors provide information in a standardized way. The extra effort required to implement this could be measured with a small number of volunteers before making it a requirement for data deposit, to both check for feasibility and ensure that the benefits of considering and tracking this information outweigh the costs.

Further acknowledgments

We would like to thank the curators of the corpora who replied to our email: Airi Kapanen, Alan Crutenden, Alison Henry, Aliyah Morgenstern, Amy Strekas, Amye Warren-Leubecker, Ana Isabel Ojea Lopez, Ana Lúcia Santos, Ana Maria Guimarães, Andra Kütt, Andrea Biró, Andrea Feldman, Angela Grimm, Ann Peters, Anna Chromá, Anna Theakston, Anne Van Kleeck, Anne-Marie Schaerlaekens, Annick De Houwer, Annick DeHouwer, 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, Carrie Dyck, Catherine Snow, Cécile De Cat, Charles Watkins, Chiara Roggero, Chien-ju Chang, Christian Champaud, Christiane von Stutterheim, Christina Gildersleeve-Neumann, Christine Howe, Christophe Parisse, Claartje Levelt, Claudine Hammerlath, Colleen Huebner Morisset, Conxita Lleo, Conxita Lleó, Cornelia Hamann, Darinka Anđelković, David Dickinson, David Gil, DMITAR Popov, Dominique Bassano,



The figure consists of two panels. The left panel is a screenshot of a software interface for entering speaker-level metadata. It includes fields for Speaker IDs (GEO), Language (eng), Corpus name (Snyder), Name code (GEO), Age (3;6.22), Sex (Male), Group (Urban), Social economic (Lower), Role (Participant), Education, Unique file ID, and Optional speaker name (George). The right panel is a text-based list of corpus-level metadata for the CHILDES Nungon Sarvasy Corpus, including Creator (Sarvasy, Hannah), Subject (child language development), Subject.olac:linguistic-field (language_acquisition), Subject.olac:language (yuw), Subject.childes:participant (age="2;11 - 4;10"), Description (children learning Nungon (Yao) in Northeast New Guinea), Publisher (TalkBank), Contributor, Date (2017), Type (Text), Type.olac:linguistic-type (primary_text), Type.olac:discourse-type (dialogue), IMDI_Genre (discourse), IMDI_Interactivity (interactive), IMDI_PlanningType (spontaneous), IMDI_Involvement (non-elicited), IMDI_SocialContext (family), IMDI_EventStructure (conversation), IMDI_Task (unspecified), IMDI_Modalities (Spoken), IMDI_Subject (unspecified), IMDI_EthnicGroup (unspecified), IMDI_RecordingConditions (unspecified), IMDI_AccessAvailability (open access), IMDI_Continent (Oceania), IMDI_Country (Papua-NewGuinea), IMDI_ProjectDescription, and IMDI_MediaFileDescription.

Figure 1: Metadata found in CHILDES corpora. The left panel shows a selection of speaker-level metadata fields (reproduced from p. 35 of the CHAT Manual, MacWhinney, 2000). The variable Group can contain any grouping relevant to the corpus producers. The right panel shows corpus-level metadata for Sarvasy (2017).

Donella Antelmi, Donna Jackson-Maldonado, Donna Thal , Dorit Ravid, Eithne Guilfoyle, Ekaterina Pro-
tassova, Elena Lieven, Elena Nicoladis, Elena Pizzuto, Elena Tribushinina, Elena V. M. 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 Wij-
nen, 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, Huang
Yue-Yuan, Hye-Ree Ghim, Igor Žagar, Iliana Reyes, Inge Zink, Ioana Goga, Isabelle Barrière, Isabelle Mail-
lochon, Jacqueline Sachs, Jacqueline van Kampen, Jan Edwards, Jane Herbert, Jane S. Tsay, Janet Bang,
Jasmina Moskovljević Popović, Javier Aguado Orea, Jean Berko Gleason, Jean Quigley, Jean-Adolphe Ron-
dal, 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, Juana Licerias, Judit Navracics, Julian Pine, Julie Brittain, Julie McMillan, Jürgen Weissenborn,
Kaja Kohler, Karina Hess Zimmermann, Karne Beek, Katerina Palasis, Katherine Demuth, Katherine Nel-
son, Kathy Post, Keith Sawyer, Kim Plunkett, Klára Matiasovitsová, Klaus Wagner, L. Haggerty, Laetitia
de Almeida, Larisa Avram, Larry F. Guthrie, Leonor Sciliar-Cabral, Liliana Tolchinsky, Linda Kelly, Linhui
Li, LinHui Li, Lise Menn, Livia Tonelli, Lois Bloom, Lori Van Houten, Lorraine McCune, Luigi Rizzi, Lynn
S. Bliss, Madalena Cruz-Ferreira, Madeleine Leveillé, Magda Krupa-Kwiatkowska, Magdalena Smoczyn-
ska, Maigi Vija, Maja Roch, Manuela Wagner, Mara Steinberg Lowe, Marc Bornstein, Margaret Deuchar,
Marguerite Mackenzie, 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 Vih-
man, 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 Red-
ford, 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, Nada Ševa, Nan Bernstein Ratner, Naomi Hamasaki, Naomi
Yamaguchi, Natalia Gagarina, Neil Smith, Neiloufar Family, Nicola Botting, Nina Gram Garmann, Norio
Naka, Not Found, 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, Sinead McNally, Sirli Zupping, Sophie Kern, Sotaro Kita, Stan Kuczaj, Stephanie Durrleman, Stephen Matthews, Steven Gillis, Sudaporn Luksaneeyanawin, Susan Ellis Weismer, Susan Gelman, Susan R. Braunwald, Susana Correia, Susana López Ornat, Susanne Miyata, Sven Strömquist, Takeo Ishii, Tamirand De Lisser, Tania Ionin, Teresa da Costa, Thea Cameron-Faulkner, Thomas Doukas, Thomas Lee, Tina Hickey, Tina Ringstad, Twila Tardif, Ulrich Frauenfelder, Ur Shlonsky, Uri Tadmor, 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, and Zhang Yibin.

Additional analyses

The following figures on education use WDI's proportion of the population completing high school (rather than Our world in data's proportion of the population completing lower secondary school).

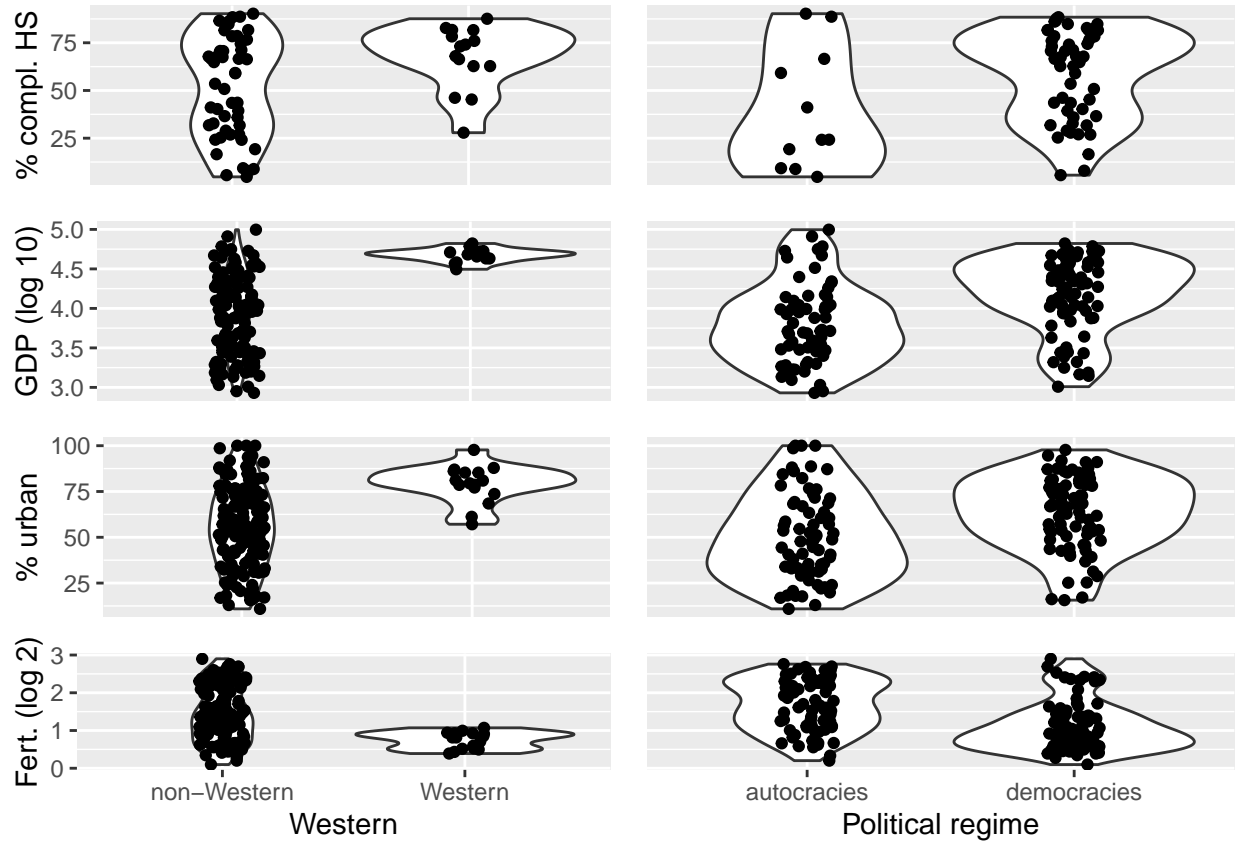


Figure 2: Equivalent to Figure 1 in the main manuscript, only the education variable has changed.

Association across vars: all countries

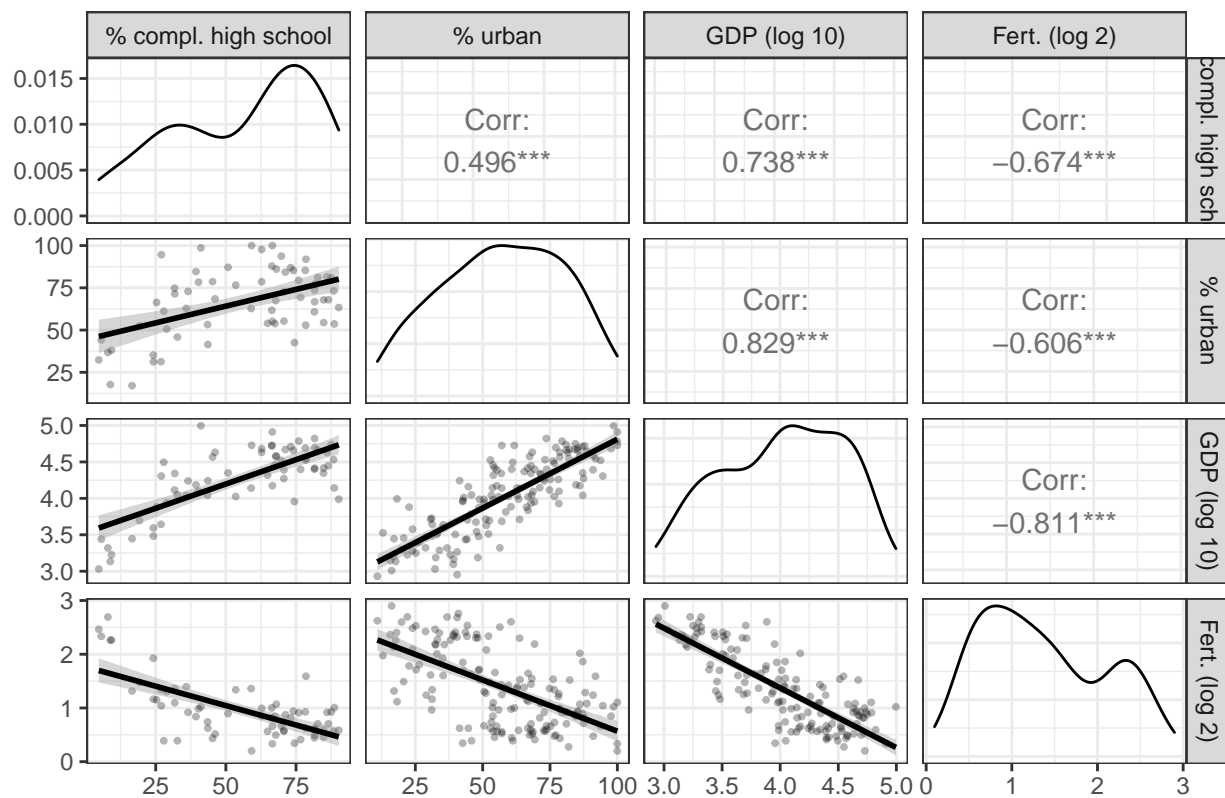


Figure 3: Equivalent to Figure 2 in the main manuscript, only the education variable has changed.

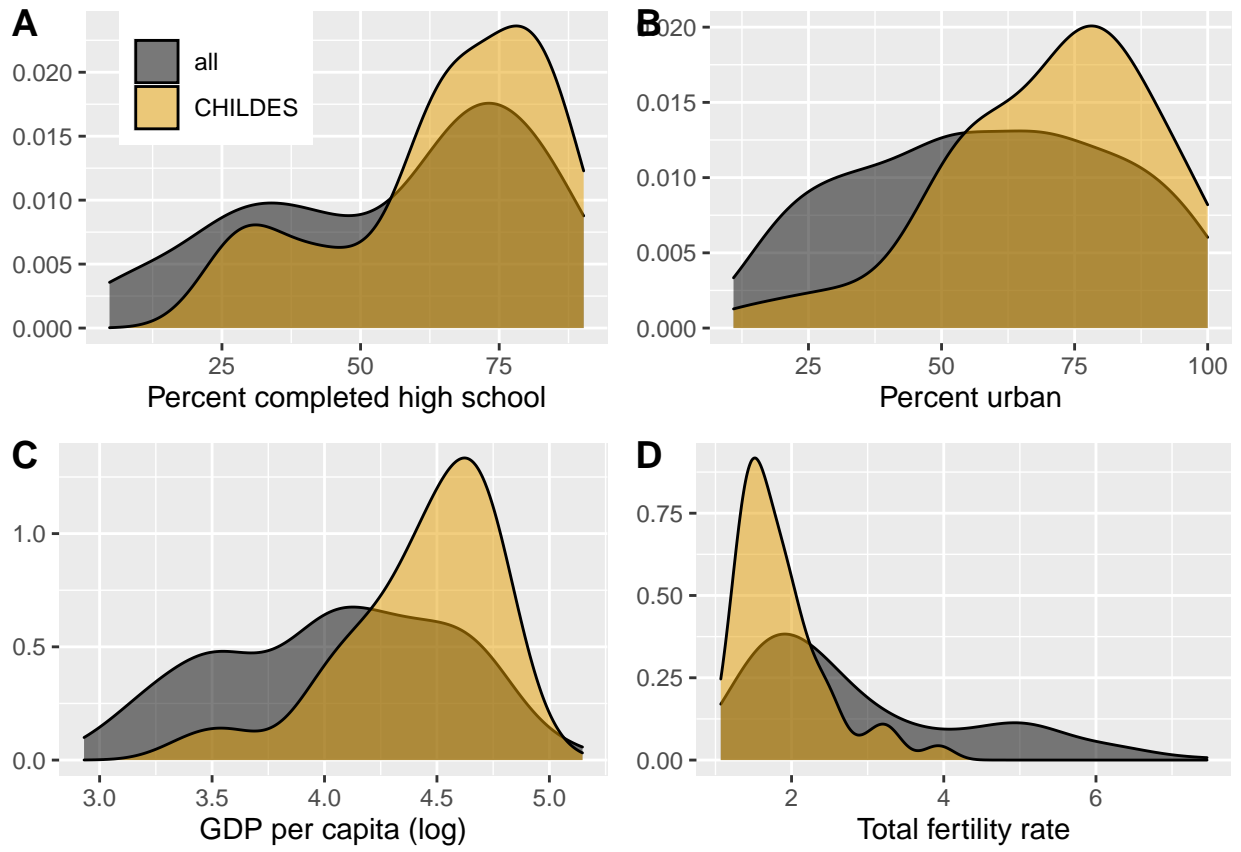


Figure 4: Equivalent to Figure 3 in the main manuscript, only the education variable has changed.