Unweaving WEIRD Patterns in CHILDES: Biases in Naturalistic Recordings

Camila Scaff, Georgia Loukatou, Alejandrina Cristia & Naomi Havron

University of Zurich, Laboratoire de Sciences Cognitives et de Psycholinguistique, ENS, EHESS, CNRS, PSL University,University of Haifa

Recent investigations into mainstream journals highlighted a significant bias in developmental studies, with a predominant focus on WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations, raising concerns about the reliability and generalizability of research findings. To further explore this issue, we focus on the CHILDES (Child Language Data Exchange System) database, which serves as the primary repository for naturalistic language recordings and transcripts used in language acquisition research. We systematically review the database to uncover potential biases of naturalistic language input samples. We assess corpora across four critical dimensions known to influence early language input: Socioeconomic Status (SES), Urbanization, Family Structure, and Languages. SES encompasses aspects such as education, wealth, and occupation, while Urbanization distinguishes between urban and rural settings. Family Structure delves into variables like the average number of children per household and the family's composition (extended or nuclear). Languages consider the number of languages in each corpus and the children's lingual status (monolingual, bilingual, or multilingual). We describe results at both country and corpus-level. While our examination of 180 corpora within CHILDES reveals a rich representation of languages and countries, it also uncovers biases across all dimensions and levels of analyses. Specifically, we find that CHILDES corpora predominantly sample from middle to higher SES backgrounds. In nearly half of the corpora, children had parents with research-related professions. Furthermore, the data is skewed towards urban settings and nuclear family structures. While naturalistic samples in CHILDES include more than 43 different languages, two-thirds of corpora feature monolingual children. In summary, the naturalistic recordings found in CHILDES predominantly feature wealthier, highly educated, urban, and monolingual nuclear families. Although these biases may impact specific research questions differently, it is crucial to acknowledge their existence. Our findings underscore the need for greater inclusivity and diversity in sampling methods within child language studies.