DATA MANAGEMENT PLAN

Raw Audio Data and Meta/Derivative Data

Given that the raw audio data for this project is intrinsically identifiable data (i.e. recordings of day-to-day activities from young children who wore audio recorders), extra steps are taken to ensure that these raw audio files are only available to authorized researchers, as specified by the data guardians, based on families' consent as regulated through each relevant institution's Research Ethics Board.

All members of the research teams working with raw audio data will undergo mandatory ethics training as dictated by their host institution (e.g. CITI Certificates in the US, CORE for Canada). Indeed, most PIs in our group already stipulate this lab-internally, and are well-versed with the infrastructure for compliance. All researchers on this project will be or are authorized HomeBank researchers; HomeBank has an approval process in place to facilitate ethical raw-audio and metadata access. The Argentina and LuCiD Corpora (not presently on HomeBank) will be shared over secure-server connections (sftp with unique logins) before being packaged into the virtual machines (see below).

Raw audio data will be stored on encrypted, password-protected machines kept in locked offices; each primary data guardian retains a copy of raw data in the home lab, ensuring mirrored archiving. All manual annotations will be backed up nightly over secured network connections maintained by Duke University IT (***Bergelson*** Lab).

Github, Virtual Machine, and OSF storage of Code, and Documentation

All code for each tool, and the manual and tool-derived annotations will be packaged into a 'child language module' via virtual machine platform, using the already-existing infrastructure developed by the Virtual Speech Kitchen (***Metze***). This will allow standardized computing environments across labs, and facilitate bug-fixes and version-control. All code will be shared both group-internally and with the research community, and will be written in either free or standalone formats (e.g. python, R).

As the tools team (***Dupoux, Metze, Räsänen, Rudzicz, Schuller***) improves their code, as the datasets team provides further annotations (***Bergelson, Soderstrom, Rosemberg, Cristia***), and as the group analyzes human- and machine-annotation for publication (**all PIs)**, each will "push" their work to a shared private github repository, to be "pulled" by the other groups. This is for intermediary control of training and testing data, and analyses, before code is ready for public use and feedback. At each "code release" landmark, and at the end of the project, code and annotations will be released publically through the project github page, linked to PI, lab, and HomeBank github code repositories for long-term storage (Spring, Fall 2018, Fall 2019, Spring 2020, see Appendix 3 & PMDC). Documentation of the developed annotation process will be maintained, stored and released longterm via the Open Science Framework.

Long-term Storage of Data

Once they reach acceptable levels for distribution, both hand-coded and automated annotations will be stored in the long term together with the original raw audio (i.e. in HomeBank and at LuCiD, in addition to the individual laboratories) so that these codes can be used for subsequent analyses by other researchers. A short-term moratorium on use of these derivative data (specified in HomeBank's dataset-specific fair-use policies) may be imposed until the main research findings have been published.

Results Dissemination

As described in the Data Dissemination section of the PMDC, results, workflows, and data will be made available to the research community and the general public via project blog, wiki, conference presentations, and open-access publications. We will take extra efforts beyond the requirements of publication to make our analysis code available via R Markdown scripts coupled with sharing of summarized de-identified data on the project website, so that other researchers can replicate our analyses directly, and apply them to their own data, which may not be sharable outright.

Our group is committed to open-source pipelines, for increased dissemination, replicability, reuse, and extension.