Developmental Language Datasets and Tools Match-Up Bootcamp

On May 22–24th, 2019 the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands, will host a tool development bootcamp during which we will pair up researchers from two groups:

Language development researchers who have very large datasets (i.e., for which complete manual annotation is not feasible) that are partially annotated for one or more language-related phenomena.
Speech technology, natural language processing, and machine learning researchers who are developing and improving tools that create automated annotations of related phenomena.

Our goal is to launch the next wave of developmental speech tools forward for a handful of tasks that are fundamental for measuring child language development and challenging for the speech tools community.

Our purpose.
Child language researchers need to to be able to measure what children hear and say in their everyday lives to effectively formulate and test theories about their language development. Recent advances in recording technology are increasingly allowing researchers to document whole waking days at a time, both in well-studied (e.g., English) and understudied (e.g., Tsimane') speech communities. Yet software for mining usable information from these recordings has lagged woefully behind, leaving most researchers to choose between expensive, out-of-date software or fully manual annotation. For tools researchers, these same datasets present an immense challenge: is their tool able to work under truly naturalistic conditions (e.g., multiple speakers in multiple speech environments) and with a population with much a huge potential for applied development (e.g., caregivers tracking their children's language development). In particular, the development of such tools for under-resourced languages is critically needed to advance developmental language study and the applicability of speech technology to new populations.

Tools and datasets.
We have already invited researchers relevant to four tools, which we will summarize below:

S*peech/Voice Activity Detection and broad speaker ID classification:* This is the single biggest bottleneck to current child language research using automated annotation. We have gathered curators of 9 databases with relevant annotations. In each case, children 0–5 years were wearing the recording device and are surrounded by a variable number of people over the course of multiple hours. The languages represented so far include English (North American and UK), Dutch, Quechua, Tseltal, Tsimane', and Yélî Dnye and the target children recorded include both typically developing participants and some at risk for autism. Cumulatively, the annotated recordings sum to at least 30h (probably much more).

Child vocal type classification: We can study children's vocal production, even before their first words, by assessing the different types of coos, babbles, and cries they use early on. This tool works as follows would classify children's vocalizations into four basic categories to estimate overall vocal maturity. We have gathered 4 database curators for this task, with the recorded children ranging in age from 0 to 3 years, both typically developing and at risk for autism. These data are primarily recorded in American English, but we have access also to data in Tseltal and Yélî Dnye.

Detecting book reading activity: Book reading has been shown to expose young children to words they otherwise don't hear and to elicit rich interaction between children and their caregivers. For that reason, many child language interventions focus on reading books. This tool would detect book-reading sessions from daylong (4–16-hour) recordings and could be used in intervention studies. We have so far only identified one volunteer with book-reading data, but their dataset includes samples from both American English- and Spanish-speaking families varying in socio-economic status.

Automatic speech recognition, spoken term detection/keyword spotting: Past work on child language development has demonstrated the immense value of tracking specific words children hear and say over the course of early development (e.g., vocabulary size is correlated with general language development). This tool would help identify specific words chosen a priori as relevant to the research question from running speech in child-centered audio recordings. We have so far found two datasets for this task including both (partly transcribed, but not word-aligned) auto recordings and secondary measures of children's vocabulary. Both of these corpora are in English.

We believe that the goals of our bootcamp are in-line with those of SIGUL, and we would love to demonstrate our connection to the SIGUL community with an official SIGUL label. Please don't hesitate to contact us with your questions and/or comments: Marisa Casillas (marisa.casillas@mpi.nl), Alex Cristia (alecristia@gmail.com), and Caroline Rowland (caroline.rowland@mpi.nl). See https://www.mpi.nl/events/ddtmatch-up/ for updates.