Supplementary materials for 'Language exposure predicts children's phonetic patterning: Evidence from language shift', by Margaret Cychosz. *Language* 98(3).461–509, 2022.

Bilingual language exposure and use predict children's phonetic patterning: Evidence from language shift: Supplementary Materials I

Margaret Cychosz

Processing daylong recordings

Selecting recordings

Because most children completed multiple daylong recordings, the first step in data analysis was to select one recording per child. To get the best estimate of each child's environment, the research team decided to annotate the longest duration recording for each child. In one case, it was not possible to use the longest duration recording because the child went to sleep with the recorder on, resulting in almost all of the recording occurring when the child was asleep. In that case, the second-longest recording was used. Most recordings selected for annotation were made with the LENA digital language processor. Two children's annotated recordings were instead made with the Zoom H1n Handy recorder either because the child only completed one day of recording and it was done with the Zoom recorder (n=1) or because the child's LENA recording contained the child sleeping overnight (n=1). The average duration of the daylong recordings used for bilingual language estimation was 12.12 hours (range 7.63-16 hours), with no notable durational outliers within any age group (see Table 3 in manuscript).

Sampling recordings

First, a custom set of Python scripts was written to process the recordings. Since families were permitted to pause the recording, pieces of the recording were stitched together to create one longer recording. There were, on average, 2.12 recording "pieces" that made up each recording (range: 1-4). The stitched pieces were interspersed with a 1000ms clip of white noise to mark boundaries between recording pieces. After stitching the recording pieces together, the entire recording was then chopped into 30-second clips. The 30-second clip length was chosen because previous work employing daylong recordings has annotated 30-second clips from recordings to estimate characteristics of children's language exposure, including bilingual language exposure (Micheletti et al., 2020; Orena, Byers-Heinlein, & Polka, 2019; Ramírez-Esparza, García-Sierra, & Kuhl, 2014) and the random sampling annotation technique was validated on 30-second clips (Cychosz, Villanueva, & Weisleder, accepted).

A standard vocal activity detector (Usoltsev, 2015) run over all of the 30-second clips reported what percentage of each clip contained vocal activity. Clips that contained 0% vocal activity were not drawn for annotation. Additionally, before beginning annotation, the author listened to portions of the recordings with low reported vocal activity to determine if the child was napping. The target child was considered to be napping if there was relative quiet in the background, no speech from the target child, and heavy breathing or snoring. Clips where the child was determined to be sleeping were marked not to be drawn for annotation. Finally, clips where the researcher was present (for example if the recorder was turned on as the researcher was leaving a participant's home) were also identified prior to annotation and were drawn but not annotated.

Corpus annotation

Recordings were annotated using a custom Generalized User Interface (GUI) application. (Now available open-source to use on additional annotation tasks at `https://github.com/megseekosh/Categorize_app_v2`.) The GUI application would randomly select a clip, with replacement, from a given participants' clips. The researcher would listen to the drawn clip and categorize the speaker(s) and language(s) heard. Research personnel had the option to repeat the clip as many times as they would like. For each clip, annotators made the following decisions:

1. **Language?**: Quechua, Spanish, Mixed, No speech, Personal Identifying Information (PID), Researcher Present, or Unsure
2. **Speaker?**: Target Child, Target Child & Adult, Other Child, Other Child & Adult, Adult, or Unsure
3. **Media Present?**: Yes or No

If there was no speech in the clip, annotators selected 'No speech' under 'Language' and moved on to the next clip. For the language choice, research personnel were instructed to identify the language being spoken in the clip. If only Spanish was spoken (regardless of the quantity of

speech), the researcher marked 'Spanish.' Similarly, the researcher marked 'Quechua' for monolingual Quechua clips. If the research personnel heard both Quechua and Spanish in the clip—code-switching within a sentence or two different conversations—they marked the clip as 'Mixed.' For those clips where the speaker or language was not clear,

For the speaker annotation, the 'Target Child' was the child wearing the recorder and 'Other Child' was defined as any individual whose voice sounded as though they had not gone through puberty. Usually the team could determine whether a speaker was a child or an adult because they had information on the household members and their ages. However, in the cases where an annotator could not recognize a voice, the team labeled the speaker as a child if their voice sounded pre-pubescent. Personnel were instructed to annotate 'Target Child and Adult,' if a clip contained the target child, another child, and an adult. See `https://github.com/megseekosh/Categorize_app_v2/blob/master/FAQs_bilingual.MD` for further details on annotation decisions, including a list of frequently asked questions used to standardize annotation between research personnel.

When the language or speaker in a clip was unclear (e.g. a caregiver singing nonce words, other non-language vocalizations), research personnel could select 'Unsure.' Language and speaker were coded separately so annotators could still code for speaker or language even if the other category was unclear. The 'Unsure' annotation was most often used for clips where a conversation was taking place in the background of the recording which made it difficult to determine the language and/or speaker. The team of annotators considered the possibility that it may be difficult to ascertain the speaker or language in some clips because those clips are noisier and contain multiple interlocutors. These clips could also be more likely to occur outside of the home. These noisy clips, with multiple interlocutors, might be more likely to contain mixed speech. Thus, disregarding these clips *could* lead the team to inadvertently disregard clips of a certain category (i.e. Mixed speech). In practice, however, the 'Unsure' clips almost always contained background speech without a discernible speaker or language, so the team felt confident in excluding 'Unsure' clips from further analysis.

The choice for media was binary—'Present' or 'Absent'—because (1) it was often difficult to determine if the media in the recording was radio or TV and (2) almost all of the media was in Spanish, making it irrelevant to mark the language. In other words, when media was present, it was in Spanish.

The team defined personal identifying information (PID) as any clip containing first *and* last names (but not just first), street addresses and specific neighborhoods (but not just the name of the town), birth dates, and any discussion of financial information. The 'Researcher Present' annotation indicated that the researcher could be heard on the recording. If PID or Researcher Present were selected for the language choice, the annotator did not continue to annotate for speaker and moved on to the next clip.

As annotators drew and listened to the 30-second clips, they were simultaneously running a Jupyter notebook (included in the Github repository) to mark progress towards annotation. The notebook recorded the proportion of Quechua, Spanish, and Mixed clips to total clips for each child. Human annotation was cut off when two criteria were met. First, the *proportion* and *variance* (Variance measured over a moving window of 60 language proportion estimates.) between language categories had to asymptote (exemplified in Figures 1, 2). Second, 50 language clips from each child had to be annotated (Language clips include those annotated as Quechua, Spanish, or Mixed, but not Unsure or No Speech.). The 50-clip criterion was included as an additional, precautionary measure, to ensure sufficient transcription even if stability between language category proportion and variance was reached. Given these pre-determined criteria, the team was more confident that their annotations were accurately reflecting the child's language environment.

Research personnel

The annotation personnel underwent a stringent training procedure prior to and during annotation. First, personnel spent approximately three hours practicing annotating 30-second clips from a single recording from the corpus (the recording was from a child who was not a participant in the current study). After this initial practice, research personnel had to pass an

Figure 1. Example area plot of language proportions by number of clips annotated. Area plots were used to track progress towards language proportion stability during daylong recording annotation.

annotation test. The lead researcher selected and annotated 40 30-second clips from the same practice recording. The lead researcher's clip annotations were considered the gold standard annotations (the lead researcher knows the families and conducted the fieldwork in Bolivia). The clips were selected to represent an array of situations that the research assistants would eventually encounter in the recordings (e.g. no speech, overlapping speech, presence of multiple interlocutors of different ages). The research assistants coded the test clips for Speaker, Language, and Media. The assistants' annotations were compared to the gold standard annotations. Research assistants could not begin annotating recordings for the project until they passed the annotation test with a score of 35/40 correct annotations.

Once the research assistants began annotating, weekly or bi-weekly check-ins were conducted. At these check-ins, the research personnel would discuss clips that they found difficult to annotate and a lengthy list of "Frequently Asked Annotation Questions" was constructed for the team to further standardize annotation between personnel (FAQ list included in the project's Github repository). The check-up meetings were also used to listen to clips together and discuss

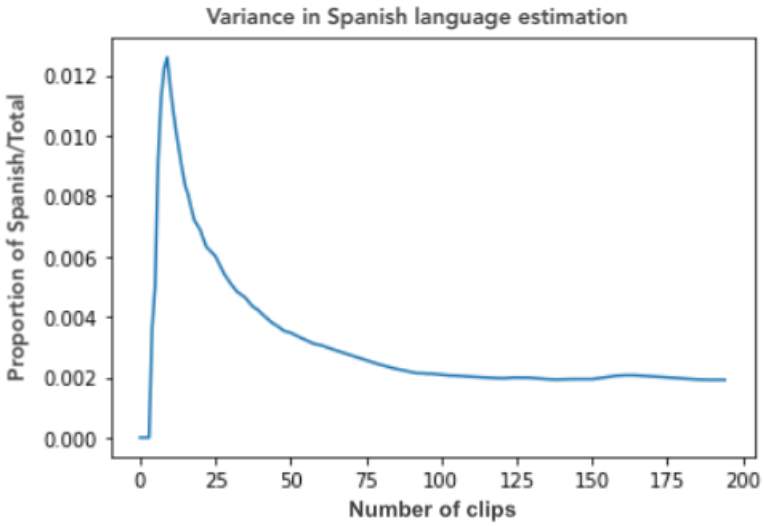

Figure 2. Example plot of Spanish language proportion variance by number of clips annotated. Variance was computed over a moving window of 60 clips. This plot was used to track progress towards variance stability during daylong recording annotation.

annotation choices to be made.

References

Cychosz, M., Villanueva, A., & Weisleder, A. (accepted). Efficient estimation of children's language exposure in two bilingual communities. *Journal of Speech, Language, and Hearing Research*.

Micheletti, M., de Barbaro, K., Fellows, M. D., Hixon, J. G., Slatcher, R. B., & Pennebaker, J. W. (2020). Optimal sampling strategies for characterizing behavior and affect from ambulatory audio recordings. *Journal of Family Psychology*. doi: 10.1037/fam0000654

Orena, A. J., Byers-Heinlein, K., & Polka, L. (2019). Reliability of the Language Environment Analysis (LENA) in French-English Bilingual Speech. *Journal of Speech Language and Hearing Research*, *67*(2), 2491–2500. doi: 10.31234/osf.io/3xcvu

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, *17*(6), 880–891. doi: 10.1111/desc.12172

Usoltsev, A. (2015). *Voice Activity Detector-Python*. (https://github.com/marsbroshok/VAD-python)