PROJECT SUMMARY

Overview:

We propose to create a public, permanent, extensible database of ecologically valid, daylong audio recordings of child and family auditory environments. The sensitive nature and large size of these recordings has inhibited the sharing of these recordings through existing mechanisms. The length of the recordings also makes automated analysis methods crucial. A system for sharing analysis code with other developers and with child development researchers is needed to facilitate the development and sharing of new automated analysis methods. The proposed project will establish three new resources for the child development and automatic speech recognition communities: (1) a public dataset containing about 1,200 hours of daylong audio recordings that have had private information removed, (2) a larger dataset containing about 12,000 to 120,000 hours of daylong audio recordings that have not had private information removed and that will be restricted to researchers who have agreed to keep the data private and have obtained training in ethical issues of working with data from human subjects, and (3) an open-source repository for code designed to automatically analyze the daylong audio recordings or to facilitate their human transcription. The shared datasets will represent a range of different groups, including typically developing and clinical groups, at a range of ages from newborn infants to preschool or even school age children, and from a range of language and socioeconomic backgrounds.

Intellectual Merit :

The ability to observe child utterances and parent-child interactions over the course of entire days in completely naturalistic settings is enabling unprecedented statistical power and ecologically validity and is transforming the study of child language development. The audio databases and the code repository will enable more scientists to access the data, promote larger scale meta-analyses, and foster interdisciplinary collaboration. The two fields we expect to benefit most from these data will be (1) child development researchers interested in language acquisition or parent-child interaction and (2) engineers focusing on the development of automatic speech and vocalization processing technologies. The audio datasets and code repository will facilitate interdisciplinary collaboration by making child language data more available to engineers and by making new audio processing methods more accessible to child development researchers. Computational modelers of child development are also expected to be among the users of the shared daylong child audio data. These tools will accelerate the speed of discovery of the roles that various factors play in language, emotional, and cognitive development. They will also accelerate the development of new technologies for processing audio data, especially in noisy environments and containing child speech, notorious for the difficulty they pose for current automatic speech recognition tools.

Broader Impacts :

Daylong audio recordings, and automated analyses of those recordings, are now being used in several large-scale early interventions with low-SES families in major U.S. metropolitan areas. By improving the data on which these early interventions are based, and by improving the tools available for providing automated parent feedback, the corpora and code repository will likely have benefits for future intervention programs. The public corpus and the open source code repository will be available for use not only by researchers but also by students. The fact that the datasets will be freely available will enable access by those who lack the financial and other resources necessary to collect their own such samples. Undergraduate, Masters, and Ph.D. students will assist with various aspects of the project. Women and underrepresented minorities will be well represented among the students involved, and a diverse group of participants will be represented in the shared datasets.

PROJECT DESCRIPTION

Overview

We propose to create a public, permanent, extensible database of ecologically valid, daylong audio recordings of child and family auditory environments. In the past 5 years, such recordings have had a major impact both in the scientific literature and in application to early childhood interventions. However, their sensitive nature and large size have inhibited the sharing of these recordings through existing mechanisms. The proposed project will create a database system called HomeBank to permanently and freely share daylong child audio files while ensuring appropriate privacy safeguards, and will seed the database with a variety of recordings from different research labs, making it immediately useful to researchers.

The length of the recordings also precludes purely manual data coding and makes the development of automated analysis methods crucial. A system for sharing code with other developers and with non-programmers is urgently needed. Developing a database and code repository for daylong child audio recordings will enable large scale meta-analyses, allow a larger number of scientists to work with the data, improve child speech recognition tools, and foster interdisciplinary collaboration. The project will establish a well-documented repository of open source code, called HomeBankCode, for analyzing daylong child audio recordings.

Background

Serious research on child speech and vocal development began in the middle of the 20th Century (Chomsky, 1959; Piaget, n.d.; Skinner, 1957), gaining widespread attention at least as early as the early 1970s (Brown, 1973; Lenneberg, 1967). The overwhelming majority of this research depended on interactions between parents and children collected in a laboratory and coded by trained transcribers. This method is expensive, time-consuming, and prone to human transcriber error and bias. It has ecological validity concerns, and often relies on small samples. Using technology developed in the past 10 years, daylong audio recordings can be readily collected in the child's natural environment, providing large, ecologically valid samples of data, with consistent formats across laboratories. Machine-learning algorithms designed to operate on these data provide automated counts of certain aspects of the child's productions and the child's linguistic environment in a tiny fraction of the time required for a human to transcribe such data. Although modern automatic methods are not free from their own weaknesses—validity of segmentation and a priori label assignments chief among them—it has opened new perspectives on child speech and vocal development, family dynamics, and assessment and intervention.

The most prominent system for daylong audio of children and families in their natural, usually home, environments is currently the LENA^{TM} system, first developed in the mid-2000's by Infoture, Inc., now the LENA Research Foundation (LRF). The LENA Pro system is in use by over 200 universities, hospitals, and other research institutions (http://www.lenafoundation.org/lena-pro/). Another well-known system for gathering daylong home audio recordings is the Human Speechome Project, which started as a case study of a single child (Roy, Frank, & Roy, 2009, 2012). To date, the Human Speechome Project has not garnered nearly as many users, and is not as readily available for researchers. Thus, the present project focuses on recordings made using the LENA system, with the aim of eventually accommodating other systems for daylong recording of home environments as they become readily available.

The LENA (Language ENvironment Analysis) system. The LENA system consists of both hardware and software components (see Figure 1). On the hardware end, the system includes a small audio recorder capable of storing up to 16 hours of audio. There is a single microphone that records unprocessed acoustic data to onboard solid-state memory. The system also includes a variety of custom-tailored clothing options, including vests, t-shirts, and overalls, each of which has a pocket sewn into the front chest into which the recorder is placed. The clothing has been tested to minimize noise from rustling of clothing. The system is robust to spills and is comfortable, safe, and easy for parents to use. In the typical usage, a researcher gives the recorder to the parent along with instructions to turn the recorder on when their child wakes up in the morning and off when the child goes to sleep at night. Unlike most wireless

recording setups used in laboratory settings, the recorder does not transmit to a receiver; instead it stores the entire recording on the device worn by the child. This allows the recorder to be worn throughout the day, regardless of the location of the child—it can even record during car rides, trips to the park, etc. Parents are told that they can pause the recorder should they need to for privacy, or entire recordings can be discarded at their request. Upon return, each recorder's data can be uploaded to a lab computer using the LENA Pro software. The recording is then processed using the software's automatic labeling algorithms. After each recording is fully processed, the data can be exported to produce a fully accessible audio file (in lossless 16 kHz, 16-bit, PCM, WAV format) and the complete results of the automatic labeling procedure (as a centi-second resolution XML formatted record of events, time-aligned to the WAV file). Summaries of the recording are also presented in the LENA Pro software's graphical user interface.

The goal of automatic speech recognition (ASR) is to obtain a transcription from the acoustic signal. Typically the goal is to map acoustic speech to readable text, but this approach has not been entirely successful in natural (i.e., ecologically valid) settings with highly variable vocalizations, such as within family and child speech. Thus the LENA system focuses on a simpler, though still challenging, automatic labeling task: that of categorizing segments labeled by the primary sound source, including talker identity (*child wearing the recorder*, *any other child*, *adult female*, *adult male*, and *overlap*), other sound sources (*TV or other electronic sounds, noise*), and silence. Using these automatically determined labels, the software also estimates additional key elements within the child and adult vocalizations, such as the number (although not the identity) of words spoken by adults and speech-related child vocalizations (e.g., speaking, babble, singing) vs. non-speech-related child vocalizations (e.g. crying, laughing, burping). Using the automatically determined speaker labels, the system also estimates when the child wearing the recorder is engaged in conversational turns with an adult speaker. Finally, the system provides an Automatic Vocalization Assessment (Richards, Gilkerson, Paul, & Xu, 2008) that aims to provide information about the child's developmental level with regard to speech production ability, based on established norms from the Natural Language Study (Gilkerson & Richards, 2008). The system relies on standard ASR methods, using a combination of Gaussian mixture models and hidden Markov models to obtain the speaker labels and the child speech-related versus not-speech-related labels (Richards et al., 2008; Xu, Yapanel, Gray, Gilkerson, et al., 2008; Xu, Yapanel, & Gray, 2009). It uses the open source Sphinx software as inputs in estimating number of adult words and in producing the child vocalization assessment (Bořil et al., 2014; Xu, Yapanel, Gray, & Baer, 2008). Data on the reliability of the LENA labels and adult word counts compared to human coders has been published for typically-developing (TD) children learning English (Soderstrom & Wittebolle, 2013; Xu et al., 2009) and Spanish (Weisleder & Fernald, 2013).

Research utilizing automated analysis of daylong audio recordings using the LENA system is advancing our understanding of autism spectrum disorders (Dykstra et al., 2013; Oller et al., 2010; Warlaumont, Richards, Gilkerson, & Oller, 2014; Warren et al., 2010), childhood hearing loss (VanDam et al., 2015; VanDam, Moeller, & Tomblin, 2010), the consequences of premature birth (Caskey, Stephens, Tucker, & Vohr, 2011; Johnson, Caskey, Rand, Tucker, & Vohr, 2014), the impact of television viewing (Ambrose, VanDam, & Moeller, 2014; Aragon & Yoshinaga-Itano, 2012; Christakis et al., 2009; Zimmerman et al., 2009), and more. These studies using the LENA system have each utilized thousands of hours of audio data.

Figure 1. Top: Participants wearing LENA recorders in previous studies conducted by PI Warlaumont. Bottom: Overview of the LENA system, which includes recorders, clothing, and software to upload and automatically label the recordings (from http://www.lenafoundation.org/).

Research on this scale would not be feasible if it relied on human transcribers.

The technology is also used in applied settings, for example in the Providence Talks (Hodson, 2014), Project Aspire (Sacks et al., 2014), and the Thirty Million Words (Leffel & Suskind, 2013) initiatives. In these projects, the LENA system is being deployed to examine the effects of poverty, hearing loss, and rehabilitation efficacy on young children's developing linguistic systems. These research projects and intervention initiatives leverage the strengths of daylong naturalistic recording combined with automatic speech processing to inform questions of interest to a wide range of groups, including medical professionals, early childhood educators, policy makers, and politicians.

Research labs around the world are thus beginning to acquire datasets of naturalistic daylong audio that are vastly larger and more representative than what was available even a decade ago. Unfortunately, privacy concerns due to the home-based nature of the recordings combined with their length have made vetting of the recordings and removal of private information particularly labor-intensive. This results in most of the collected data remaining sequestered in individual labs. **The development of a system for sharing these rich data collections would confer a great benefit to the fields of child language acquisition and automatic speech recognition. Such a system would support the sharing and development of resources for basic research as well as for educational and clinical work. Facilitating the development of a system for sharing this valuable data as well as improvements on and extensions to the existing analysis tools are the key motivations for the proposed project.**

CHILDES (The Child Language Data Exchange System) and TalkBank. CHILDES is a web-based system for sharing and analyzing child language transcript data. The database at childes.talkbank.org includes 50 million words of transcript data, much of it linked on the sentence level to digitized video or audio recordings that can be played back directly over the web. Over 95% of the data in CHILDES are publicly available for downloading and analysis without a password. CHILDES is one component of the larger TalkBank system (talkbank.org) that includes additional databases for the study of aphasia, traumatic brain injury, second language acquisition, dementia, conversation analysis, and other language areas.

CHILDES began in 1984 with support from the MacArthur Foundation and has received continuous support from NIH since 1987 and from NSF between 1999 and 2004. There are currently 1,800 users of CHILDES located in 35 countries. A search at scholar.google.com reveals 5,482 articles in English that have made use of CHILDES data or programs. However, because this inventory does not include research papers published in other major languages, the actual size of the research literature generated by CHILDES is closer to 6,800 publications. Publications based on CHILDES touch on every major issue in child language, from phonology to intellectual development. The data are most heavily used by researchers in Linguistics, Psychology, Computational Linguistics, Speech and Hearing, Sociology, and Modern Languages.

CHILDES and the more general TalkBank system of which it is a component (Figure 2) have adopted rigorous international standards for data preservation, documentation, and access. In recognition of this, TalkBank has received the Data Seal of Approval, based on accurate adherence to a set of 16 standards regarding corpus documentation (childes.talkbank.org/manuals/), consistent data formatting in a tightly controlled XML schema (talkbank.org/softare/xsddoc/), metadata generation in the OLAC (www.language-archives.org) and CMDI (clarin.eu) formats, articulation of a full mission statement, IRB protection (talkbank.org/share), data storage, long-term preservation, OAI harvesting of CMDI and OLAC metadata, PID (persistent digital object identification) through the Handle System (handle.net), backup systems (mirror sites, archiving, git, etc.), statement of codes of conduct (talkbank.org/share/ethics.html), and proper treatment of copyright (CC BY-NC-SA 3.0). TalkBank and CHILDES are also members of the international CLARIN consortium of national language data centers (clarin.eu) and MacWhinney is the Chair of the Scientific Advisory Board for CLARIN.

In addition to these achievements as a stable center for the sharing of language data, TalkBank has developed standards, programs, and practices that make it ideal as development site for the HomeBank system this project will create. For those segments of HomeBank that will be transcribed, CHILDES provides comprehensive software for analysis of phonological, morphological, syntactic, and discourse

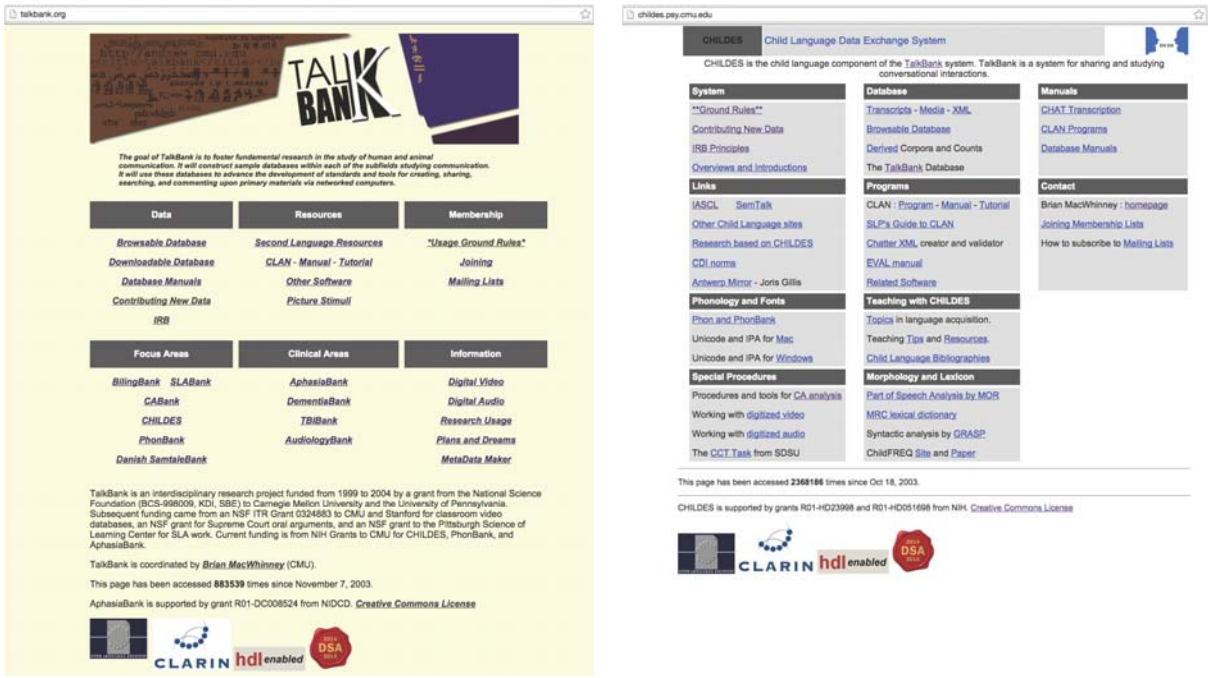

Figure 2. The TalkBank (left) and CHILDES (right) homepages. Members will be able to log into the website at http://talkbank.org/HomeBank to browse and download data, analyses, and programs free of charge.

features, much of it fully automated. These programs include CLAN for transcript analysis and Phon for phonological analysis, both with linkages to Praat, a free acoustic analysis software. CHILDES has 30 years of experience in setting up alternative levels of data access (talkbank.org/share/irb/options.html) that protect individual privacy in accord with high-level IRB standards. CMU provides 1TB ethernet access to the CHILDES servers which currently make 5TB of data available online with additional material available on request.

The HomeBank database will require a major extension of data storage and will introduce some new file types, unique privacy considerations, and a greater emphasis on raw audio and dependence on automatic transcription tools. Nevertheless, much of the needed infrastructure is within the current scope of the CHILDES/TalkBank system. The Speech Technology group in the Language Technologies Institute at Carnegie Mellon (in which MacWhinney is an associated faculty member) can provide additional state-of-the-art methods for automatic processing of HomeBank records.

Objectives

To maximize the value of daylong home audio recordings of children and their families, a system for sharing the recordings across research groups and with students, clinicians, and other groups is needed, as is a system for sharing tools used to facilitate efficient and accurate analysis of such recordings. Sharing daylong home audio presents a number of unique challenges compared to other speech corpora; foremost among them are the substantial privacy concerns and the infeasibility of human transcription. The proposed project thus has three main objectives:

1. Create a public database of ***vetted*** daylong audio recordings.
2. Create a larger, restricted database of ***unvetted*** daylong audio recordings.
3. Create an open source code repository for processing daylong audio recordings.

The products of Objectives 1 and 2 will constitute the new HomeBank system and the product of Objective 3 will constitute the new HomeBankCode repository.

Objective 1: Create a public database of VETTED daylong audio recordings

The first objective of the proposed project is to create a *public* database containing recordings made with the LENA system. Sharing as much data and resources as possible in a very public way has a number of advantages, particularly in increasing accessibility of the data to a wide range of students, researchers, and parents. The data will be protected through a password that will be given to all qualified researchers, students, and interested parents, teachers, and clinicians, who will accept terms of agreement specifying how they are permitted to repost clips of the data; the data will thus be essentially public to humans, but difficult for non-human bots to access. Not needing an extensive application process in order to access the data will make the process of utilizing the data more efficient both for the users and the managers of HomeBank. A number of researchers collecting daylong child audio recordings are already asking parents to specify how broadly the data can be shared. In the proposed project, families will be asked to choose from these three options: (1) data will only be available within the lab collecting the data, (2) data can be shared with other researchers, or (3) data can be made totally public (see Table 1).

Table 1. Summary of Data Sharing Levels.

	Who may access	Main advantages	Est. # of recordings in HomeBank after 3 years
Level 1	Restricted to a single lab and its direct collaborators.	• Highest level of participant privacy. • Appropriate for particularly vulnerable participant groups and participants not comfortable with data sharing. • HomeBankCode programs can still be used for data processing and analysis.	N/A
Level 2	Available to all HomeBank members with IRB approval who have signed the usage agreement promising to maintain data confidentially.	• Private and sensitive information is restricted to those trained to work with such data and committed to protecting participants. • Analyses developed using transcribed datasets can be run on a larger sample size. • Good for unsupervised machine learning algorithms requiring large amounts of unlabeled data. • Subsets can be transcribed.	1,000–10,000 (~12,000–120,000 hours)
Level 3	Available to all interested individuals who create a HomeBank member account and consent to the usage agreement.	• Broad, easy-to-obtain access to researchers, students, parents, and clinicians. • Private and sensitive information are removed. • A subset will be transcribed. • Transcribed portions provide training data for supervised machine learning algorithms. • Suitable for pilot research or when a very large dataset is not required.	100 (~1200 hours)

PI Warlaumont at the University of California, Merced and Consultant Bergelson at the University of Rochester have already obtained IRB approval for sharing LENA data with other researchers and, at UC

Merced, with the general public, provided that last names are removed. Many families in ongoing daylong audio recording studies (all with institutional and IRB oversight) have already given permission for their data to be shared publicly at Level 3, and others are willing to share data at Level 2. Obtaining IRB approval for data sharing of daylong LENA audio is feasible.

Participant Consent. We will construct consent form templates that allow parents to opt in to sharing their data in HomeBank. Permission will when possible be sought at the time of each recording. However, there will also be times, especially early in the project, when it is necessary to seek permission retrospectively. We will solicit advice from our team consultants when constructing the consent forms. The forms will be approved by our own IRBs and by the IRBs at all contributing researchers' institutions.

The consent forms will ask parents to indicate whether they approve for their child's daylong audio recordings and metadata to be included in a shared database and whether they are comfortable with the public being able to listen to their recordings or whether they would want their data only in the restricted dataset, for download by more thoroughly vetted researchers only. If they consent for data to be shared publicly, the form will also ask them which kinds of data they would want removed, and will offer suggestions such as last names, addresses, and arguments as well as an opportunity to write in other types of events. The consent forms will also provide contact information should they decide at a later date that they wish to revoke access to their data, with clarification that shared data may have been downloaded by users prior to access revocation by the family. Parents will also be offered the opportunity to listen to recordings or segments of recordings themselves before deciding whether to approve their inclusion in the public dataset.

Vetting and removal of private segments. Even when parents are willing to share their child's daylong audio recordings publicly, there are still privacy concerns that should be addressed before making those recordings public. The primary issue is that parents and other individuals in the child's environment can easily forget or be otherwise unaware (a service person in the home, for example) that they are being recorded. In those cases, it is possible that parents or others being recorded will reveal information that should not be released publicly. For instance, parents may have an embarrassing argument, or reveal identifying personal information such as addresses, birthdates, or last names. Thus, research assistants trained to flag sensitive information will listen in full to any recording that is to be made public. Any segments containing sensitive information will be removed prior to making the recordings publicly available in HomeBank. Numerous tools are available for performing these deletions.

All students and research personnel performing the vetting will be IRB-approved to work with sensitive data. The research programs, laboratory spaces, and equipment of all principal contributors have extensive experience working with child and family data, and safeguards in excess of federal and state minimums are strictly maintained.

Metadata. Demographic information and other metadata will also be included in the database. Metadata will include child age, child sex, information about whether the child is typically developing or is a member of a specific clinical population (e.g., having hearing impairment, language delay, autism spectrum disorder, etc.), education level of the child's primary caregivers, country the child was recorded in, dominant ambient language, scores on language or other developmental tests or questionnaires, and comments on the recording. The metadata protocol will be designed to be flexible and extensible within the database. For example, certain characteristics of children with autism spectrum disorder may warrant a number of data fields unique to that population. As with the determination of which information should be flagged for deletion from recordings, we will use standard practices and advice from our team of expert consultants. All decisions will also be governed by active institutional oversight as described above.

Automatically labeled events. Certain events of interest will also be included in the database. The LENA system's automatic labeling routines identify segment boundaries for the vocal activity of specific talkers (adult female, adult male, target child, etc.) and environmental acoustics including noise, overlapping vocals, silence, and presence of television or radio audio. Using the automatically generated labels, conversational exchanges are defined as temporally close alternation of vocal activity between two speakers. These conversational exchanges can be examined for frequency, contributions by specific

participant types (adult males, usually corresponding to fathers, for example), or other features. The automatic labels are all available within an XML file called an ITS file that can be exported by the LENA software. ITS files will be included in the database and linked to the audio and metadata for the same recording. Apart from the date of birth, the other information provided in an ITS file is not identifiable. There will be cases where ITS files (with date of birth info. removed) can be shared althought the audio data cannot. Summary data can be included in the metadata for each recording.

Human transcriptions. While automatic labeling is extremely valuable because it is computed without supervision and can be efficiently applied to the entire dataset, it has a number of limitations. First, the labels are not as accurate as those that can be made by human listeners. Second, the labels only provide very basic information about the events in the recording. Many researchers will be interested in not only when individuals are vocalizing but also in the phonetic, prosodic, linguistic, or semantic *content* of those vocalizations. Thus, transcriptions will also be included in the databases. PI VanDam will manage a team of speech-language pathology Masters students who will provide orthographic and broad phonetic transcriptions of a substantial portion of the public recordings in CHAT and Phon formats, the inter-convertible formats currently used for TalkBank data. Volunteer undergraduate research assistants in PI Warlaumont's lab can provide additional assistance with transcription, particularly at the orthographic level. The trained research assistants will provide transcriptions of approximately 100 hours of recordings, providing an extremely useful supplement to the larger database. The automatic labeling software can be used to focus transcription on portions of recordings that have certain properties, for example, high child vocal activity. The 100 hour estimate is based on an estimate that on average it will take approximately 30 hours of time transcribe an hour of audio recording.

The transcription files will also be included and linked to the other files representing a given recording. The transcriptions will have a number of potential uses. First, they will allow for direct comparison of the linguistic characteristics of child speech in the ultra naturalistic contexts represented in the daylong home audio recordings. One might expect the behaviors of children and family members captured in daylong home recordings to differ compared to those captured in shorter, more targeted recordings of parent-child interactions in home and laboratory contexts. Thus, the transcriptions of HomeBank data and the transcriptions of CHILDES data will be complementary rather than redundant. Second, the transcriptions will serve as training data for automatic voice and speech processing algorithms developed specifically for daylong audio recording data (see Objective 3).

Given the long duration of the recordings, it would take too long to transcribe all the daylong audio recordings. We will encourage HomeBank users to transcribe the dataset for their own research purposes and to contribute their transcriptions to the database at the time of publication.

Distribution and preservation. The public database will be preserved, backed up, and hosted/distributed via TalkBank (http://talkbank.org/).

Objective 2: Create a larger, restricted database of UNVETTED daylong audio recordings

Because of the time and labor involved in preparing recordings for public access, there will be many more recordings for which parents have given permission for their data to be shared with other researchers but for which we do not have the resources to listen and clean up the data to share the data publicly. In this sense, the data truly represent "big data". Despite being unlabeled, this large volume of unvetted audio recordings has huge potential to be useful for studying child development and for developing new voice and speech recognition technologies. Thus, HomeBank will accommodate not only vetted data, but also unvetted data, to be made available in a more restricted manner (Level 2 access).

Potential use by speech processing engineers. The raw WAV files included in the unvetted dataset will provide an amply large dataset for input to unsupervised machine learning systems. For example, unsupervised deep learning neural network algorithms are increasingly being recognized as powerful methods for extracting acoustic features for automatic speech analysis (Hinton et al., 2012). In this approach, raw audio input is first transformed into a basic set of acoustic features, which are then input to an artificial neural network that funnels the representation of the waveform into progressively smaller neural layers. This has the effect of compressing the audio, i.e. representing it with a smaller number of

features than the raw audio. The neural activity in the small layers is then expanded out to progressively larger and larger layers, culminating in an output layer that is the same dimensionality as the original input layer. The neural network connection weights are optimized to maximize the match between the input data and the reconstructed output for a training set of data. The more training data that is available, and the more representative it is of the full range of sounds that one can be expected to encounter in the speech recognition application, the better the acoustic features will be.

The lower-dimensionality acoustic features, namely those in the smaller, internal layers of the neural network, can then be used as input to supervised machine learning algorithms that learn to associate acoustic features with classes of sounds within the audio, such as phonetic categories, speaker labels, and so on. The transcriptions of Level 3 HomeBank data could be used as labels for training supervised learning systems and for evaluation of automatic speech processing systems. Furthermore, the availability of the unvetted audio recordings to approved groups of listeners would enable anyone with the human resources necessary to contribute to providing additional labels that could be used for training their own systems and that would, ideally be shared back to the rest of the community using the recordings. In essence, providing the full set of unvetted, daylong audio recordings for use by approved users would provide valuable data to speech processing engineers. This would presumably help them generalize their methods to anohtr type of naturalistic child data. The development of better speech processing methods trained on naturalistic speech would in turn benefit the child development community, because better automated analysis tools will enable more efficient and higher quality research and assessment.

Potential use by child development researchers. The data in the unvetted corpus will also provide an important resource to child development researchers. In contrast to in-lab experiments or short home recordings, daylong recordings of children and their caregivers in their natural environments provide more holistic information about child development. They provide a window into all types of children's daily experiences, as well as the ability to better estimate the frequency of different types of events and experiences over the course of a child's day. Many researchers, including student researchers, may have questions that can be addressed by human or automated coding of these daylong naturalistic samples, but lack the time or funds to obtain a large number of LENA recorders themselves. Even if they do have the time and funds, those resources are not best spent on replicating work that has already been conducted by another laboratory but on collecting data that enables increases in sample size or the addition of new demographic or clinical groups.

Once users are approved to use the unvetted Level 2 portion of HomeBank, they will be able to access the full audio recordings and their meta-data. For instance users may train IRB-approved listeners working under their supervision to find segments of the recording in which events of a particular type (e.g., singing or book reading) are present and manually transcribe those events, or perform acoustic analyses on them. Alternatively, users may apply their own automated data analysis methods to a large number of the unvetted recordings without necessarily listening to the recordings except to establish reliability. One advantage of daylong recording is that, assuming there is a way to efficiently process the large quantity of data, it is possible to accumulate data on relatively rare events. In both cases, this arrangement ensures that hard-earned daylong recording data, generously contributed by parents and children for the purposes of increasing our understanding of child development, is put to maximum use.

Obtaining approval. Because the data in this larger, unvetted Level 2 repository will often contain private episodes which should not be shared with the broad public, researchers will need to undergo a more rigorous approval process before they can access this part of HomeBank. In particular, we will require that researchers demonstrate that they have CITI training, IRB approval from their own institution consistent with the original IRB approvals under which the recordings were obtained, and understanding of and commitment to the HomeBank Level 2 user agreement. The user agreement will specify very prominently that the recordings are to be kept on password-protected machines in locked rooms at their institutions or on their laptops provided that they are encrypted. The agreement will also require training of all students or other individuals acting under the investigators' supervision who interact with the dataset, to ensure that all who interact with the data are aware of the private nature of the data and committed to maintaining privacy of the participants represented in it. The specifics of the user agreement

will distinguish between Level 2 HomeBank users who desire access to unvetted material only for technological processing without direct listening to the content of the recordings from users who will be performing content-related processing.

We will model this approval system on two previous systems: the AphasiaBank corpus in TalkBank for sharing language samples and brain imaging data from patients with aphasia and the Databrary system for sharing video data from child development studies. PI MacWhinney was one of the creators of AphasiaBank and is highly experienced with managing such restricted datasets. Additionally, we have been in discussion with the Databrary team, and they have agreed to consult for our project. *Our goal will be to make the data as easily accessible as possible while maintaining the appropriate safeguards needed to ensure participant privacy.*

Distribution of the data. Because the data included in the unvetted corpus are expected to be much larger than those in the vetted and partially transcribed corpus, it will take longer to for users to download the data. We will implement a data transfer mechanism that allows for individuals to resume data transfer if the user needs to pause the transfer or becomes disconnected. Transfers will be both password protected and encrypted. In the case that the dataset grows to be on the order of tens of terabytes, users who wish to use the entire dataset as opposed to subsamples of it can be sent the data on encrypted, password protected hard drives through a tracked and insured mail service. Users will be responsible for paying the shipping costs, and will be able to inspect samples of the full dataset beforehand to ensure it meets their needs. This process can also be used in instances where individuals do not have access to fast or reliable Internet connections. Data and applications to access the restricted dataset will be distributed and submitted primarily via the TalkBank website.

Vetting and transcribing the unvetted dataset. When other researchers utilize the unvetted dataset, in many cases, their work will involve listening to and transcribing portions of the data. We will require approved researchers to log periods where private data are present, using the same guidelines as we will use in creating the public, vetted dataset. This way the data may over time be shared more publicly. As with the public dataset, we will also urge the researchers to make their transcriptions using an already established standard format such as CHAT and require them to share their transcriptions according to pre-established guidelines (for example, when they have published papers using their transcriptions of the dataset). In this way, we can increase the size of the public dataset (subject to participant permission for their data to be shared publicly, and not only with other researchers) as well as increase the quantity of data for which there are human labels for events of interest. Such labeled events are a likely bottleneck in utilizing the data to develop and validate new automatic speech processing technologies. Having users help contribute human transcriptions for segments of the unvetted restricted library will help to ameliorate this bottleneck, at little cost to the researcher and with the added benefit of promoting more open and replicable science.

Objective 3: Create an open source code repository for processing daylong audio recordings

A number of research groups have already begun to develop extensions to the automated analyses provided by the LENA Pro software. Below we first describe what some of these extensions accomplish, then we describe our plan for building an open source code repository called HomeBankCode. HomeBankCode will make the extensions more readily available to potential users and accelerate the development of new extensions.

Acoustics of child and adult vocalizations. Several extensions focus on acoustic analyses to get more detailed pictures of the sounds the children and adults are producing. For example, Oller et al. (2010) analyzed child vocalizations by first segmenting them on the basis of amplitude contours into vocal islands, which roughly corresponded to syllables, and then analyzing those vocal islands according to duration, spectral tilt, and various other acoustic features. They then put these vocal island level acoustic features through principal components analysis and used the principal components as inputs to classifiers of clinical group membership. They found that the approach could reliably discriminate the recordings of children with autism from those of children with developmental delays but not autism and from those of typically developing children. Unfortunately, the computer programs from these analyses

have not yet been made publicly available. Even so, the paper was one of the most downloaded papers in *PNAS* in 2010. Making the code for extensions like this publicly available will increase the impact on future research and clinical practice.

PIs VanDam and Warlaumont and Consultants Soderstrom, Seidl, and Cristia have been combining the output of LENA with acoustic analysis tools available in Praat (Boersma & Weenink, 2015) and custom routines in MATLAB, Python, and R to obtain acoustic information (pitch, vowel formants, spectral characteristics, amplitude, etc.) for child, female adult, and male adult vocalization segments.

One example of a current direction of this work is characterizing the acoustics of child-directed speech (CDS) and adult-directed speech (ADS) during the ultra-naturalistic interactions represented in the home recordings. Characteristic CDS has increased pitch, extended syllable and word durations, exaggerated prosody, restricted syntax, and greater phonetic variability (Broen, 1972; Fernald et al., 1989; Hoff-Ginsberg, 1985; Snow & Ferguson, 1977). Although CDS in general has had steady attention in the literature at least since the early 1970s, CDS is receiving intense attention recently due to technical advances and a renewed appreciation for ecological validity that allow for new perspectives on the topic. For example, some studies have found that fathers and mothers differ in the speech they direct toward their children (Mannle & Tomasello, 1987; Reese & Fivush, 1993; Tenenbaum & Leaper, 1997, 1998, 2003). This possibility has been examined recently using a very large database of LENA recordings, showing that in daylong, ecologically valid samples, compared to mothers, fathers used fewer pitch fluctuations (VanDam & De Palma, in press), a greater variety of lexical items, and more complex syntactic forms. It is possible that fathers (subconsciously) supply a "bridge" between the language spoken in the home and that used in public (Gleason, 1975).

Characterizing child-adult interaction dynamics. At a higher temporal level of analysis, PI Warlaumont and her collaborators have developed tools that use the onset and offset times of child and adult vocalization segments as identified by the LENA Pro software to give a richer picture of the overall pattern of when children and adults are vocalizing over the course of the day and how the children and adult's vocalizations relate to each other temporally (Warlaumont et al., 2010, 2014; Abney, Warlaumont, Haussman, Ross, & Wallot, 2014). This work has, for example, found that adult vocal responses are more likely when child vocalizations are speech-related than when they are not speech related and that a child is more likely to produce a vocalization that is speech-related when the child's most recent speech-related vocalization received a response (Warlaumont et al., 2014). Furthermore, various components of this feedback loop were found to differ for children of different ages and socioeconomic backgrounds as well as for children who are typically developing compared to those with autism spectrum disorder. Combined with computational modeling work (Warlaumont, 2014), the results provided support for the theory that there is a positive feedback loop between child behavior and reinforcing adult responses that helps support children's speech development, and that differences in the feedback loop can have cascading effects on the child's overall developmental trajectory (Karmiloff-Smith, 1998; Leezenbaum, Campbell, Butler, & Iverson, 2013). These results provide an example of how automatic speaker identification within daylong home audio recordings can provide the quantity of time series data needed to detect the presence of a two-part feedback loop. The code for these analyses was provided as supplemental material to the paper, and another researcher, Alex Cristia at EHESS in Paris, has since verified that the code and documentation were sufficient to fairly easily replicate the analyses. Warlaumont is currently developing extensions of this code that utilize additional statistical controls, for use by collaborators in Paul Yoder's lab at Vanderbilt University. These programs as well as other tools developed by Warlaumont and her students will be included in the shared open source code repository that will be curated as part of the proposed project.

Development of new tools for interacting with data. Other tools being developed include those that make human listening tasks more efficient. PIs Warlaumont and VanDam have collaborated informally on developing a tool for automatically extracting audio segments specified in a LENA ITS file from the associated WAV file, playing all the sounds from a given talker to the user, and allowing the user to provide feedback, for example on whether or not the sound was assigned the correct speaker label. The ability to make other judgments, for example, regarding the syllabicity or pitch of the vocalization can be

straightforwardly added. PI VanDam's research group has further developed custom software for audio extraction and playback and human listener judgment collection, acoustic analysis, database management, and statistical analysis (Ambrose et al., 2014; VanDam, Ambrose, & Moeller, 2012; VanDam & De Palma, in press).

A major advantage to this specialized playback software is that it enables much more efficient human coding of the data. VanDam & Silbert (2013) reported on human judgments of over 90,000 exemplars of speech segment labeled by the automatic methods of the LENA system, and accuracy/validity was compared for different label types. It was found that the automatic methods perform at a similar to other state-of-the-art ASR methods, but that performance varies for different labels (adult men are more accurately labeled than adult women, for example). Acoustic analyses showed that temporal and spectral qualities—but not amplitude—interact in complex ways in the automatic label determination process.

This data extraction, playback, and acoustic analysis software has been explicitly developed in a format suitable for sharing, extensibility, and modification. Early versions of VanDam's and Warlaumont's programs have been shared since 2010 via the LENA User Forum, a discussion forum administered by the LENA Research Foundation. Although a number of professional users of the LENA system have expressed interest in using the programs, the adoption of the extensions by other researchers has been limited by the lack of a centralized repository where up-to-date versions can be accessed and by a lack of high quality documentation.

Sharing these Extensions and their Code. In order to better share these LENA extensions across research labs, we will build an open source code repository called HomeBankCode. A GitHub Organization account will be created, allowing not only for multiple contributors but also for multiple administrators, making the repository truly a community resource. GitHub provides free storage and version history. It also has the advantage of being extremely widely used across academia, industry, and hobbyists, making it likely that many potential contributors are already familiar with how to use Git and GitHub and increasing the chances that other users will discover the resource (Figure **3**).

Creating a place where developers can share code will not in itself be a difficult task. The greater challenges will be in (1) advertising the resource to other developers and providing resources so that they can easily incorporate it into their workflow and (2) making the code in the repository usable to a wide audience of researchers, including those with limited computer programming skills. Since we will address the first challenge later in the project description, we will focus here on our plan to address the second challenge.

Our first task will be to put our own LENA analysis extensions into the shared repository and document them so that they are usable by other researchers, including those without advanced programming skills. After a beta of the documentation is created, we will test the software and documentation with undergraduate and Masters students within our own labs, thereafter expanding it to the growing network of LENA-using researchers who can apply it to their own datasets. We will work as closely as needed with these researchers and their students to enable them to utilize the code.

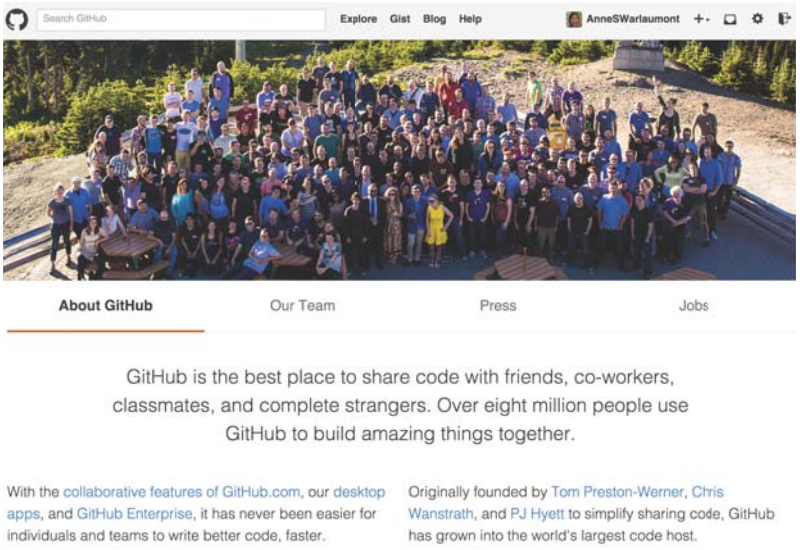

Figure 3. Screenshot from the GitHub website. https://github.com/about

The arrangement will be mutually beneficial as they will receive hands-on assistance tailored to their needs, and we will receive feedback on which aspects of the software and/or the documentation are in need of improvement. Presumably the application of the tools to new datasets will in itself lead to new findings and publications.

After making our own software accessible via the repository, in consultation with other repository contributors and LENA users, we will create a set of suggested best practices for other code contributors to make their code maximally accessible and interoperable with other programs. For example, when feasible and when it does not create too much burden on the developer, developers will be encouraged to use common conventions in specifying inputs and outputs of their programs. We will also offer hands-on assistance during the funding period to developers interested in contributing code to the repository, helping them with documentation and with making minor modifications to increase ease of use and interoperability. A wiki for HomeBankCode developers will be used to disseminate this information.

An Example of How the Three Tools Could Be Put to Use

In this section we provide an example of a project that could be supported by the proposed infrastructure, demonstrating the roles that the three tools proposed above could play in the research and development process. A topic of much discussion and excitement among LENA researchers is how to develop tools to automatically categorize adult speech as child-directed or other-directed. One of the most exciting findings using the LENA system is that quantity of child-directed (but not adult-directed) speech predicts child vocabulary growth (Weisleder & Fernald, 2013). In recent conference calls with a "LENA Meets ASR" network coordinated by consultant Alex Cristia, the community has determined that one of our highest priorities should be developing a program that can automatically label points during a LENA recording when an adult is directing speech toward a child, as opposed to toward other individuals. Such a tool would be useful not only to researchers but also in applied settings for providing reliable feedback to caregivers of at-risk children.

Such a project would first use the public, vetted (Level 3) HomeBank dataset of daylong home audio recordings (as described in Objective 1) to create a dataset where segments of LENA recordings are linked to transcriptions of when adult vocalization segments are child-directed and when they are not.

Once a dataset of LENA recordings transcribed according to child-directedness is available, computational research groups can develop automated algorithms for classifying recording segments according to child-directedness using the frequencies and temporal patterns of the LENA Pro's speaker ID labels, the acoustics of the vocalizations, or a combination of the two. This algorithm could be trained and validated using the human listeners' gold-standard child-directedness transcriptions.

In undertaking this work, human transcribers would make use of tools available in the HomeBankCode repository for efficient listening to and labeling of the LENA recording segments, while the automated algorithm might make use of tools in the repository for characterizing patterns of child-adult interactivity and/or for obtaining acoustic information about vocalizations (see Objective 3).

Once the automated algorithm is validated on the human transcribed data, it could be applied to the larger set of recordings available in the restricted, unvetted (Level 2) HomeBank dataset, in order to estimate how much child-directed speech children at different ages, from different cultural and socioeconomic backgrounds, and/or from different clinical and control populations are exposed to (see Objective 2). Thereafter, human listeners could transcribe subsets of the unvetted database according to child-directedness of adult speech, in order to test the reliability of the automated algorithm, and perhaps also according to child words produced in order to estimate the relationship between child-directedness of adult speech on vocabulary development.

Once the study is published, several research products would be contributed back to the toolset, including the *transcriptions* that were made of the subsample of unvetted audio data and the *software and algorithms* for automatically identifying child-directed adult speech.

Creating Awareness Among the Scientific Community

During the funding period, it will be critical to create ample awareness of and enthusiasm for the new child data and code repository resources we will be developing. We will take a number of steps to promote HomeBank and HomeBankCode among relevant scientific communities. We will present and give hands-on workshops at international conferences, including meetings of the Society for Research on Child Development, the International Association for the Study of Child Language, the Cognitive Development Society, and Interspeech. PI and consultant meetings will sometimes be co-located with these meetings, reducing travel time and costs.

We will also make periodic announcements about the development of the tools and the addition of new datasets and software tools on listserves, such as the info-childes and cogdevsoc mailing lists. We will also reach out personally to specific researchers who have published studies using the LENA system and who have developed new speech technologies, to invite them to consider contributing data or code and to consider using the public databases and software in their future studies.

We will build a website that describes the resources being developed as part of this project and provides information for prospective data contributors, contributors of analysis code, and potential users of any of the resources. The UC Merced undergraduate summer students and the project coordinator will help with designing and updating this website, and will work in coordination with the other project members and consultants to make it informative and easy to navigate. The website will describe each of the project components as well as how they relate to each other and to other existing resources and will direct visitors to where the audio recordings and computer programs can be accessed.

Finally, we will publish the results of our own work using the public and restricted databases and using the open source software from the repository in peer-reviewed journals, including at least one open-access journal. The UC Merced Ph.D. student researcher will take the lead on writing up at least one such paper during the second and third years of the project. Publishing an open-access paper that describes utilization of the tools to serve scientific aims will generate awareness of the tool and provide a concrete example of how the resources can be used.

Attributing credit to data and software contributors. It will be very important for researchers who generously contribute their data and/or code to be duly acknowledged for their efforts. Following the model of the CHILDES database, all users of data from either the public or the restricted datasets will be provided with the appropriate citations and required to cite them in any publications that utilize the datasets. CHILDES has placed consistent emphasis on the importance of citing original sources. As a result there has not yet been a publication in which CHILDES data were used without citation of the original source. CHILDES also maintains methods for citing corpora as publications through assignment of ISBN codes and Handle System IDs; this will be applied to HomeBank data as well. Code in the HomeBankCode GitHub repositories will be licensed under a Creative Commons Attribution-NonCommercial or similar license requiring the user to acknowledge the creators of the source code, unless the contributor specifies that they do not desire such acknowledgement, and prohibiting non-commercial use of the code. (Commercial use of the tools that involve the LENA system will require negotiation with the LENA Research Foundation; contact information will be provided.)

Collaboration Plan

The three PIs, Anne Warlaumont, Mark VanDam, and Brian MacWhinney, have complementary skill sets and experiences and will collaborate on all three components of the project. The three PIs, the Project Coordinator, and the Ph.D. student will hold project conference calls once per month, to assess progress, discuss issues that have arisen, and identify next steps for each group. Subgroups will meet more frequently as needed to perform the day-to-day tasks. The PIs will also convene in person twice per year.

Warlaumont will be the project leader. Her lab at UC Merced will coordinate the efforts of the various project personnel across the three locations, including arranging meetings with the consultants. Her lab will also lead the effort to recruit HomeBank data contributors and HomeBankCode developers as well as users of both the audio data in HomeBank and of the programs in the HomeBankCode repositories. Her undergraduate research assistants will perform much of the vetting of the public dataset

and will assist with the orthographic transcription effort. Her team will also lead the development of HomeBankCode. She will also advise the Ph.D. student on developing, executing, and publishing a research project utilizing the HomeBank databases and HomeBankCode repositories. Warlaumont has been working on automatic processing of prelinguistic infant vocalization data since she began her Ph.D. studies in 2007 and has been working with the LENA system since 2009. She has developed new methods for quantifying and visualizing interaction patterns captured by the system and has collected LENA recordings in both longitudinal and cross-sectional designs as well as simultaneous recordings in a daycare context. She has made several tools available publicly, including an online infant vocalization type training tool (IVICT, www.babyvoc.org), various computer programs for processing and analyzing LENA recording data (through the LENA Users forum, journal supplementary material, and her lab website), and code for computational models of vocal learning and evolution (available on GitHub). She has published research on suffix acquisition using CHILDES and CLAN and uses TalkBank in her teaching.

VanDam's role in this project focuses primarily on two aspects of the HomeBank database. First, he will coordinate and supervise the transcription efforts necessary to establish the public/vetted Level 3 portion of the database. He is skilled in transcription procedures, with more than 15 years of experience and has availability of trained students to work under his supervision, as well as excellent physical space and support services at WSU. Second, he is skilled in computer algorithms and programming, including in automatic vocalization processing, and will be a major contributor to HomeBankCode. He has extensive experience with the LENA system, being an early adopter and sustained user of the technology since 2008, and having collected over 1,200 daylong recordings to date. His research activity, including as Director of the Speech and Language Lab at WSU, is dedicated to developmental aspects of speech and language production, especially in disordered populations.

MacWhinney's role in this project focuses on adaptation of the established CHILDES/TalkBank structure to the goals of HomeBank. In terms of the technical infrastructure, he will ensure that HomeBank makes full use of all the available resources of CHILDES, including software, web integration, hardware integration, backup, sustainability, metadata, documentation, and data protection. In terms of social infrastructure, he can guarantee smooth integration of HomeBank with other data access efforts in which he participates including CLARIN, LDC (www.ldc.penn.edu), Databrary (databrary.org), DSA, OLAC, SALT, SLDR (sldr.org), LREC/ELRA, RDA (rd-alliance.org), and LinkedData (linkeddata.org). He has organized mailing lists, such as info-childes@googlegroups.com, to help promulgate awareness regarding the use of shared data. On a theoretical level, MacWhinney has studied first and second language development for 40 years, publishing 280 articles and 5 books on these topics. In 2011, he received the first Roger Brown Award from the International Association for the Study of Child Language for his work on child language learning.

Meetings with consultants will take place four times per year, with one meeting each year taking place in person. Both Warlaumont and VanDam are participants in monthly international conference calls with a group of about 2 dozen researchers called “LENA Meets ASR”, organized by Consultant Alex Cristia; the group has had three calls so far and provides an excellent example for how international conference calls within this community can be successfully executed. The consultants will provide advice on all project efforts and will in many cases be contributors to HomeBank and HomeBankCode.

Broader Impacts of the Proposed Work

Reducing Inequality in Language Acquisition. The most obvious societal impact of HomeBank will be on national discussions of issues such as the “30 Million Word Gap” (Colker, 2014; Fernald, Marchman, & Weisleder, 2013; Hart & Risley, 1995). Without comprehensive, publicly available data of the type that will be available in HomeBank, the exact nature of this purported word gap will remain obscure. Apart from such major social issues, HomeBank can contribute to the development of ASR (automatic speech recognition) methods for processing children’s speech and language in family settings. As methods progress for ASR of these materials, transcripts automatically generated from HomeBank audio will lead to an exponential increase in the available data on child language acquisition, effectively

generating extremely rich and dense corpus data on language learning. The same methods used in HomeBank with children can also be applied to the study of other language interactions, such as second language learning or language usage in aphasia.

Supporting Data Sharing. HomeBank will be accessible to researchers, students, and other interested individuals from around the world. The resource will enable access to daylong audio recording data for researchers without the resources at their own institutions to purchase the LENA Pro system and collect their own data. We will facilitate making the HomeBankCode programs and documentation as easy to use and understand as possible, making it more accessible to a diverse audience, including students. NSF and NIH have an affirmed commitment to the process of data sharing that requires that all research with human subjects include a plan for data sharing that is in accord with the requirements for protecting the privacy of human subjects. The HomeBank and HomeBankCode projects are a direct response to this important commitment. The HomeBank project will result in significant cost savings and quality improvements for the overall system for research support.

Training Students in Open Data and Interdisciplinary Research. The project will financially support students at the Ph.D., Masters, and Undergraduate levels. In addition, a number of research assistants who are either volunteers or receiving course credit will be included. Students will develop skills in data-intensive child language research and/or in computational science and computer programming that will prepare them for future careers in science, engineering, health, and education. They will gain skills in working and communicating as part of an interdisciplinary, international team. They will also gain an appreciation for data sharing and participant privacy issues.

Participation of Underrepresented Students. 62.1%, of the undergraduate students at UC Merced are first-generation college students. The campus is recognized by the U.S. Dept. of Education as a Hispanic-serving institution. PI Warlaumont's laboratory reflects this diversity. Of the Ph.D. students and undergraduate research assistants she has mentored in the past three years, 17 are women, 13 are members of underrepresented minority groups, and at least 9 are first generation college students. PI VanDam has mentored over 35 female Masters students, and, given the strong representation of female students in VanDam's academic program, it is likely that students recruited to work in VanDam's lab will be women. We are eager and prepared to mentor a diverse group of students with interests in language development and speech technology.

Direct Public Engagement. PI Warlaumont co-organizes an annual professional growth day for participants from the Merced County Office of Education and area school districts. Early childhood educators and clinicians learn about research on early language learning and network with university scientists. Graduate and undergraduate students gain experience communicating their science to a broad audience. Students involved in the proposed project will participate in this annual event. PI VanDam is a member of the Hearing Oral Program of Excellence School for hearing impaired children in Spokane. The preschool serves children with hearing loss and other disabilities. VanDam is also a member of the Riverpoint Interprofessional Education and Research group dedicated to integrated service provision to children and families. He also has ties with local medical professionals and gives regular talks at local health care providers on advances in the field of speech and language research, including talking at local and regional regular 'Grand Rounds' forums. In addition, VanDam will maintain efforts with local hospitals and continuing education of speech-language pathologists to disseminate ongoing work with this project. PI MacWhinney periodically contributes public press stories on child language issues.

Results of Prior NSF Support

None of the PIs has been the PI or Co-PI on an NSF award in the past five years. MacWhinney's development of TalkBank has received ongoing NIH support for CHILDES, AphasiaBank, and PhonBank. There has not been NSF support for TalkBank since the end of an NSF grant in 2004.

REFERENCES

Abney, D. H., Warlaumont, A. S., Haussman, A., Ross, J. M., & Wallot, S. (2014). Using nonlinear methods to quantify changes in infant limb movements and vocalizations. *Frontiers in Psychology*, *5*, 771. doi:10.3389/fpsyg.2014.00771

Ambrose, S. E., VanDam, M., & Moeller, M. P. (2014). Linguistic input, electronic media, and communication outcomes of toddlers with hearing loss. *Ear and Hearing*, *35*(2), 139–147. doi:10.1097/AUD.0b013e3182a76768

Aragon, M., & Yoshinaga-Itano, C. (2012). Using Language ENvironment Analysis to improve outcomes for children who are deaf or hard of hearing. *Seminars in Speech and Language*, *33*(4), 340–353. doi:10.1055/s-0032-1326918

Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer (Version 5.4.05). Retrieved from http://www.praat.org

Bořil, H., Zhang, Q., Ziaei, A., Hansen, J. H., Xu, D., Gilkerson, J., … others. (2014). Automatic Assessment of Language Background in Toddlers Through Phonotactic and Pitch Pattern Modeling of Short Vocalizations. In *Workshop on Child Computer Interaction (WOCCI)*. Retrieved from http://www.utd.edu/~hynek/pdfs/WOCCI14.pdf

Broen, P. A. (1972). The Verbal Environment of the Language-Learning Child. ASHA Monographs, No. 17. Retrieved from http://eric.ed.gov/?id=ED098768

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Caskey, M., Stephens, B., Tucker, R., & Vohr, B. (2011). Importance of Parent Talk on the Development of Preterm Infant Vocalizations. *Pediatrics*, *128*(5), 910–916. doi:10.1542/peds.2011-0609

Chomsky, N. (1959). A review of B. F. Skinner's Verbal Behavior. *Language*, *35*(1), 26–58.

Christakis, D. A., Gilkerson, J., Richards, J. A., Zimmerman, F. J., Garrison, M. M., Xu, D., … Yapanel, U. (2009). Audible television and decreased adult words, infant vocalizations, and conversational turns: a population-based study. *Archives of Pediatrics & Adolescent Medicine*, *163*(6), 554–558. doi:10.1001/archpediatrics.2009.61

Colker, L. J. (2014, March). The word gap: The early years make the difference. *Teaching Young Children*, *7*(3).

Dykstra, J. R., Sabatos-DeVito, M. G., Irvin, D. W., Boyd, B. A., Hume, K. A., & Odom, S. L. (2013). Using the Language Environment Analysis (LENA) system in preschool classrooms with children with autism spectrum disorders. *Autism*, *17*(5), 582–594. doi:10.1177/1362361312446206

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*(2), 234–248. doi:10.1111/desc.12019

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501.

Gilkerson, J., & Richards, J. A. (2008). *The LENA Natural Language Study* (Technical Report No. LTR-02-2). Boulder, CO: LENA Foundation. Retrieved from http://www.lenafoundation.org/wp-content/uploads/2014/10/LTR-02-2_Natural_Language_Study.pdf

Gleason, J. B. (1975). Fathers and other strangers: Men's speech to young children. In D. P. Dato (Ed.), *Developmental psycholinguistics: Theory and applications* (pp. 289–297). Washington, DC: Georgetown University Press.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore: Paul H. Brookes Publishing Co.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., … Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97. doi:10.1109/MSP.2012.2205597

Hodson, H. (2014). Automatic voice coach gives conversation tips to parents. *New Scientist*, *221*(2954), 22. doi:10.1016/S0262-4079(14)60226-8

Hoff-Ginsberg, E. (1985). Some contributions of mothers' speech to their children's syntactic growth. *Journal of Child Language*, *12*(02), 367–385. doi:10.1017/S0305000900006486

Johnson, K., Caskey, M., Rand, K., Tucker, R., & Vohr, B. (2014). Gender differences in adult-infant communication in the first months of life. *Pediatrics*, *134*(6), e1603–e1610. doi:10.1542/peds.2013-4289

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences*, *2*(10), 389–398. doi:10.1016/S1364-6613(98)01230-3

Leezenbaum, N. B., Campbell, S. B., Butler, D., & Iverson, J. M. (2013). Maternal verbal responses to communication of infants at low and heightened risk of autism. *Autism*, *18*(6), 694–703. doi:10.1177/1362361313491327

Leffel, K., & Suskind, D. (2013). Parent-directed approaches to enrich the early language environments of children living in poverty. *Seminars in Speech and Language*, *34*(4), 267–278. doi:10.1055/s-0033-1353443

Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley.

Mannle, S., & Tomasello, M. (1987). Fathers, siblings, and the bridge hypothesis. In K. E. Nelson & A. van Kleek (Eds.), *Children's language* (Vol. 6, pp. 23–42). Hillsdale, NJ: Erlbaum.

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., … Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(30), 13354–13359. doi:10.1073/pnas.1003882107

Piaget, J. (n.d.). *Logique et connaissance scientifique [Logic and scientific knowledge]*. Dijon, France: Gallimard.

Reese, E., & Fivush, R. (1993). Parental styles of talking about the past. *Developmental Psychology*, *29*(3), 596–606. doi:10.1037/0012-1649.29.3.596

Richards, J. A., Gilkerson, J., Paul, T., & Xu, D. (2008). *The LENA automatic vocalization assessment* (Technical Report No. LTR-08-1). Boulder, CO: LENA Foundation. Retrieved from http://www.lenafoundation.org/wp-content/uploads/2014/10/LTR-08-1_Automatic_Vocalization_Assessment.pdf

Roy, B. C., Frank, M. C., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*.

Roy, B. C., Frank, M. C., & Roy, D. (2012). Relating activity contexts to early word learning in dense longitudinal data. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*.

Sacks, C., Shay, S., Repplinger, L., Leffel, K. R., Sapolich, S. G., Suskind, E., … Suskind, D. (2014). Pilot testing of a parent-directed intervention (Project ASPIRE) for underserved children who are deaf or hard of hearing. *Child Language Teaching and Therapy*, *30*(1), 91–102. doi:10.1177/0265659013494873

Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Snow, C. E., & Ferguson, C. A. (Eds.). (1977). *Talking to children: Language input and acquisition*. Cambridge University Press.

Soderstrom, M., & Wittebolle, K. (2013). When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. *PLoS ONE*, *8*(11), e80646. doi:10.1371/journal.pone.0080646

Tenenbaum, H. R., & Leaper, C. (1997). Mothers' and fathers' questions to their child in Mexican-descent families: Moderators of cognitive demand during play. *Hispanic Journal of Behavioral Sciences*, *19*(3), 318–332. doi:10.1177/07399863970193005

Tenenbaum, H. R., & Leaper, C. (1998). Gender effects on Mexican-descent parents' questions and scaffolding during toy play: a sequential analysis. *First Language*, *18*(53), 129–147. doi:10.1177/014272379801805301

Tenenbaum, H. R., & Leaper, C. (2003). Parent-child conversations about science: The socialization of gender inequities? *Developmental Psychology*, *39*(1), 34–47. doi:10.1037/0012-1649.39.1.34

VanDam, M., Ambrose, S. E., & Moeller, M. P. (2012). Quantity of parental language in the home environments of hard-of-hearing 2-year-olds. *Journal of Deaf Studies and Deaf Education*, *17*(4), 402–420. doi:10.1093/deafed/ens025

VanDam, M., & De Palma, P. (in press). Fundamental frequency of child-directed speech using automatic speech recognition. In *IEEE Proceedings of the Joint 7th International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on Advanced Intelligent Systems*. Kitakyushu, Japan.

VanDam, M., Moeller, M. P., & Tomblin, B. (2010). Analyses of fundamental frequency in infants and preschoolers with hearing loss. *The Journal of the Acoustical Society of America*, *128*, 2459. doi:10.1121/1.3508806

VanDam, M., Oller, D. K., Ambrose, S. E., Gray, S., Richards, J. A., Xu, D., … Moeller, M. P. (2015). Automated vocal analysis of children with hearing loss and their typical and atypical peers. *Ear and Hearing*, 1. doi:10.1097/AUD.0000000000000138

VanDam, M., & Silbert, N. H. (2013). Precision and error of automatic speech recognition. *Proceedings of Meetings on Acoustics*, *19*(1), 060006. doi:10.1121/1.4798466

Warlaumont, A. S. (2014). An iterative probabilistic model of speech-related vocalization rate growth due to child-caregiver interaction. In *IEEE International Conference on Development and Learning and Epigenetic Robotics* (pp. 262–268). doi:10.1109/DEVLRN.2014.6982991

Warlaumont, A. S., Oller, D. K., Dale, R., Richards, J. A., Gilkerson, J., & Xu, D. (2010). Vocal interaction dynamics of children with and without autism. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 121–126). Austin, TX: Cognitive Science Society.

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A social feedback loop for speech development and its reduction in autism. *Psychological Science*, *25*(7), 1314–1324. doi:10.1177/0956797614531023

Warren, S. F., Gilkerson, J., Richards, J. A., Oller, D. K., Xu, D., Yapanel, U., & Gray, S. (2010). What automated vocal analysis reveals about the language learning environment of young children with autism. *Journal of Autism and Developmental Disorders*, *40*(5), 555–569. doi:10.1007/s10803-009-0902-5

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152. doi:10.1177/0956797613488145

Xu, D., Yapanel, U., & Gray, S. (2009). *Reliability of the LENA (TM) language environment analysis system in young children's natural home environment* (No. LTF-05-2). Boulder, CO: LENA Foundation. Retrieved from http://www.lenafoundation.org/TechReport.aspx/Reliability/LTR-05-2

Xu, D., Yapanel, U., Gray, S., & Baer, C. T. (2008). *The LENA(TM) language environment analysis system: The interpreted time segments (ITS) file* (Technical Report No. LTR-04-2). Boulder, CO: LENA Foundation. Retrieved from http://www.lenafoundation.org/TechReport.aspx/ITS_File/LTR-04-2

Xu, D., Yapanel, U., Gray, S., Gilkerson, J., Richards, J., & Hansen, J. (2008). Signal processing for young child speech language development. In *Proceedings of the 1st Workshop on Child, Computer, and Interaction*. Chania, Crete, Greece.

Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., & Yapanel, U. (2009). Teaching by listening: the importance of adult-child conversations to language development. *Pediatrics*, *124*(1), 342–349. doi:10.1542/peds.2008-2267

DATA MANAGEMENT PLAN

Audio recording data

This project focuses on creating and curating a database of extant daylong child and family recordings collected as part of previous and on-going studies. This project does not seek to collect new data, but instead to create a repository and point of access for similarly structured data from many researchers. The daylong audio recordings will, at least at first, primarily be made using the LENA system, and as such will have consistent primary and metadata file types.

UPL files. Each audio recording is originally collected as a UPL file. This is the file that is uploaded from the recorder to the user's computer. It contains the audio data as well as the unique serial number of the LENA recorder (a.k.a. digital language processor, DLP) used to make the recording and the date and time when the recording was made. A UPL file for a daylong recording is about 500MB. The UPL file serves as the input to the LENA Pro software, which, when combined with the user-specified birthdate and sex of the child who wore the recorder, subsequently performs other processing and exporting functions. UPL files are accessible via the LENA Pro software for certain functionality such as limited playback, but are not otherwise transparent to third-party applications. We will instruct data contributors on where to find these UPL files within their file systems. We will aim to obtain and store the UPL file for each recording, because inclusion of the UPL file allows for files to be reprocessed when updates to the automatic labeling software are released.

WAV files. The LENA Pro software includes a straightforward export function that converts the UPL file to a more transparently accessible WAV file. WAV is an uncompressed format that contains the audio data, plus technical information such as sampling rate and bit-depth for universal playback and analysis. WAV files are one of the most common audio formats and are completely transparent. All audio recordings made with the LENA system are single channel pulse-code modulated (PCM) audio recorded at 16 kHz sampling rate and 16-bit resolution. A WAV file for a daylong recording is about 1.5GB. Although UPL and WAV files duplicate the audio data of each recording, maintaining both files in the database allows users much greater overall flexibility: UPL files allow users to reprocess original audio with updated LENA Pro software while WAV files allow users who do not have access to the LENA Pro software to make use of the recordings. The LENA Pro software currently sells for approximately $10,000, plus $800 per year for software updates and customer support.

ITS files. When the LENA Pro software processes the UPL file, it creates an ITS file which is an XML output file summarizing the results of the automatic speech processing algorithms. ITS files are about 5MB each. XML files are ASCII-encoded text files that are human-readable but also ideally suited for automatic processing and archiving. The ITS file contains three essential sections of information about the recording. First, static characteristics of the recording and environment are included in the header. These details include recording start and stop time, recording date, the birthdate and sex of the child, the version of the recorder used to make the recording, and the version of the software used to process the recording. Second, the body of the ITS file reports the results of the automatic labeling algorithm at centi-second resolution, time-aligned with the WAV recording. This portion of the ITS file details the start and stop time of each labeled segment and the label for that segment. Labels include one of about 60 labels including *child-wearing-recorder, adult female, adult male, TV/electronic media, silence, noise*, and *overlapping vocals*. Hierarchical categories are also represented in this record, indicating combined segments that constitute a predefined *conversational exchange*, for example, in which talkers engage in specific turn-taking routines. This portion of the ITS file is typically 20,000-50,000 lines of output. Third, summary characteristics of the automated procedures are represented in the last section of the ITS output file. Details include the estimated word or syllable count of each adult vocalization, total conversational turns between the target child and adults, total number of conversations initiated or terminated by specific interlocutors, and the like. ITS files can be shared publicly, once we have removed the date of birth information, because without the date of birth they contain no identifiable information. Information about how to interpret the data in the ITS files is provided in the form of a technical report by the LENA Research Foundation, and PI Warlaumont will contribute to TalkBankCode her parser for

obtaining basic statistics regarding the information contained the ITS files and for obtaining more easily readable information about key events of interest.

Metadata. Metadata will include the child's age, sex, race, language background, caregiver education background, country, available standardized scores on language, cognitive or other developmental assessments, and known clinical information about the child (e.g. typically developing, born prematurely, hearing impairment, autism spectrum disorder, etc.). We will fashion the metadata system to match the format used for other TalkBank data.

Transcription files. Transcription data will be in CHAT and Phon formats, for consistency with the other TalkBank data.

Estimated number and size. For each daylong audio recording, the combined audio and transcription files will require approximately 2 GB of storage. We expect that the Level 3 vetted dataset will represent up 100 audio recordings once the project is complete, so about 200 GB will be required for the public storage. This quantity of data will be fairly easy to distribute via Internet. For the Level 2 restricted dataset, we anticipate between 1,000 and 10,000 recordings, requiring between 2 and 20 TB.

Downloading recordings. Recordings from the Level 3 (vetted) dataset will be downloadable via Internet from the HomeBank project website, which will be hosted on the TalkBank domain and linked from the main TalkBank homepage. Users will be able to select filtering options to download from single files to the whole dataset. Users will be provided with the option to resume download in cases they need to pause the transfer or lose Internet connectivity. As for the larger, Level 2 restricted dataset, users will first need to obtain permission to access and use the files. We will verify their completion of human subjects research training and their agreement to a data use policy before providing them with access via username and password. To remain in accord with IRB restrictions, we ask that all graduate student users work with a faculty person who will serve as the official member. Users will have the option to download subsets of the public dataset or the entire dataset. As with the public dataset, for larger transfers, a mechanism and documentation will be provided to allow for transfers to resume if the user needs to pause the transfer or loses Internet connectivity. As the database grows, or for users with poor Internet access, we will also provide an option to transfer data via encrypted and password protected hard drive shipped with tracking and insurance. Users will cover the shipping costs.

No cost to users. There will be no charge for the use of HomeBank materials or the CLAN programs.

Security. All portable machines on which private participant data is stored will be required to be encrypted and password protected. All Internet transfers will be password protected and encrypted.

Backup. All data and code associated with the project will be stored at each of the PIs' institutions, and backups performed frequently.

Revoking Access. All participating data contributors will have the opportunity to request that their data be removed from the datasets. The permission forms will note this possibility, and will also point out that previously shared data cannot be as easily revoked.

Computer Programs

All the code for the computer programs developed and revised as part of this project will be distributed publicly via a set of open GitHub repositories. GitHub allows for group administration of repositories, so that the original developer, the project personnel, and other individuals with relevant expertise can jointly manage the repositories. This will make the system more robust to personnel changes, for example as students or postdocs involved in a project move on to other positions. Copies of programs will also be stored at the PIs' campuses; storage requirements are minimal.

We will document and test the code and provide documentation, including commented code, readme files, and user manuals, on how to operate and/or modify it. Efforts will be made to ensure that they run on multiple platforms. The goal will be for programs to be accessible to individuals with limited or even no programming experience. We will encourage programs to be developed using free languages and tools such as Python, R, Perl, C/C++, and Praat but languages requiring proprietary software (e.g., MATLAB) may also be used. DOIs pointing to the relevant computer code repositories can be included in all publications that use the repositories.

SUSTAINABILITY PLAN

It is of the utmost importance to ensure that the data and code that are generously provided both by researchers and participants is maintained for use by the research community after the award expires.

HomeBank. We will guarantee the sustainability of the HomeBank data through five methods:

1. ***TalkBank Linkage***. We will link the corpora, programs and methods of HomeBank to the overall TalkBank framework. TalkBank has benefitted from 26 years of continued funding from NIH and additional funding from NSF. Current funding lasts until 2019 and the fact that funding requests for TalkBank have consistently been reviewed in the top 1% of NIH grants indicates that continued funding should be available, particularly as the project adds new interest areas.
2. ***University Commitment***. Carnegie Mellon University and the University libraries have implemented a policy of guaranteeing sustainability for the results of funded projects. This policy means that the University will continue to maintain access to the results of the project.
3. ***International Integration***. Increasingly, the survivability of datasets and methodologies is based on the adoption of international standards and linkage to global networks. HomeBank benefits from the fact that TalkBank is already an established CLARIN center, and that it has close data-sharing relations with the Linguistic Data Consortium and Databrary. If funding for HomeBank and TalkBank overall were to become inadequate, we have agreements with LDC, CLARIN, and Databrary that the data in HomeBank would be migrated to their systems.
4. ***International Access***. We will implement international control of access to and analysis of HomeBank data through the international CLARIN SPF (service provider federation) that allows trusted access to data sites for academics using systems such as InCommon within the United States and similar systems abroad.
5. ***Group Involvement***. In the context of work on CHILDES, PhonBank, AphasiaBank, and TBIBank, we have learned that the single most important factor in sustaining a growing and vital database is the integration of the database into the research agenda of the relevant community. In this case, we will work to make sure that HomeBank is directly responsive to the needs of people studying language development in the home context.

We will purchase 30–50TB data servers, 10–30TB more than the expected upper limit of what the HomeBank dataset is likely to grow to during the three funded years. This will provide ample storage for a few additional years. Minimal maintenance could be accomplished for about $12,000 per year; continued curation, extension, and outreach to the community of users will be more.

HomeBankCode. The open source code repository will be permanently maintained for free by GitHub as long as the code remains open access, so storage will incur no estimated annual operation costs beyond the funding period. Developers can continue to add their code. Our ability to support the documentation of these programs will be more limited without continued funding, however we expect to remain able to provide basic support for at least a few years past the expiration of the grant. By the end of the three years of the project we will have developed a document that outlines best practices for code documentation and for formatting inputs and outputs, which contributing developers can use for guidance. DOIs will be assigned to significant releases of the programs in the repository. In the unlikely event that GitHub becomes no longer available, we will host the repository on the TalkBank servers or other servers hosted by members of the stewardship council (they are not expected to require much space) and redirect the DOIs to those new locations.

Stewardship Council. In the final year of funding, we will assemble a group of individuals who will promise to serve for a set number of years as stewards of the HomeBank and HomeBankCode resources. The group will establish bylaws for keeping this stewardship council populated indefinitely. Members of the stewardship council will serve on a volunteer basis. Individuals who are stakeholders in the project's success, i.e. individuals who have contributed data or code to the project or whose research has utilized the project resources, will be recruited. These individuals will seek funding for hardware and personnel in order to keep HomeBank and HomeBankCode operating most effectively and to accommodate increases in its size and scope as researchers continue to contribute data and code.

TECHNICAL PLAN

Anticipated technical issues and approaches to overcoming them are presented in the Project Description and Data Management Plan.

Project milestones for each half year of the funding period, along with the primary personnel involved, are provided below. UCM = UC Merced, PI Warlaumont; WSU = Washington State University, PI VanDam; CMU = Carnegie Mellon University, PI MacWhinney.

First Half of Year 1 (Fall and Winter 2015):

- Recruit Project Coordinator. UCM.
- Purchase and set up computing equipment. UCM, WSU, CMU.
- Design data sharing permission forms; disseminate for submission to local IRBs. UCM, WSU.
- Design recording metadata dimensions and format. UCM, WSU, CMU.
- Develop procedures for vetting the public dataset. UCM, WSU, CMU.
- Hire and train Masters and Undergraduate students for transcription. WSU.
- Identify Ph.D. student with strong programming skills and good fit to research interests. UCM.
- Set up GitHub Organization Account. UCM.
- Add existing code to GitHub repository and begin documenting and testing. UCM, WSU.
- Begin setting up TalkBank to accommodate UPL, WAV, and ITS files, plus metadata and transcriptions. CMU, UCM, WSU.
- Draft project website. UCM, CMU.
- Kickoff personnel meeting hosted by UCM.
- In-person meeting and conference call with consultants.

Second Half of Year 1 (Spring and Summer 2016):

- Seed the shared datasets for beta testing with approx. 100 recordings. UCM, WSU, CMU.
- Test HomeBank data access among project personnel. UCM, CMU, WSU.
- Launch project website and announce on info-childes and cogdevsoc listserves. UCM, CMU.
- Solicit additional recordings to be included in the shared datasets. Assist with requesting consent from prior participants. UCM.
- Vet audio, stripping private material, and add to the public dataset. UCM, WSU.
- Transcribe subsets of the public audio files and add transcriptions to the public dataset. WSU.
- Select developers from other labs to upload code to HomeBankCode GitHub repositories. Help test and document. UCM.
- Two conference calls with team of consultants.
- Submit to INTERSPEECH (San Francisco, Sept. 2016) and ASHA (Philadelphia, Nov. 2016).

First Half of Year 2 (Fall and Winter 2016):

- Acquiring more recordings to be included in the shared datasets. UCM, WSU.
- Continue vetting & stripping private material for public dataset. UCM, WSU.
- Transcribe subsets of the public audio files and add transcriptions to the public dataset. WSU.
- Create best-practices guidelines for structuring and documenting HomeBankCode contributions. Begin adapting code to fit best practices. UCM, WSU.
- Apply some analysis methods from the open source repository to data in the shared databases to answer a research question. UCM.
- Recruit more broadly for HomeBank and HomeBankCode contributors. UCM.
- In-person meeting and conference call with the consultants.

- Present at INTERSPEECH (with goal of recruiting users to use HomeBank and HomeBankCode for automatic audio processing program development) and ASHA.
- Submit workshop proposals to SRCD and IASCL or arrange collocated 1-dayworkshop(s).

Second Half of Year 2 (Spring and Summer 2017):

- Continue acquiring recordings to be included in the shared datasets. UCM, WSU.
- Continue vetting & stripping private material for public dataset. UCM, WSU.
- Transcribe subsets of the public audio files and add transcriptions to the public dataset. WSU.
- Continue work applying analysis methods from the open source repository to data in the shared databases. Determine what results to submit for publication. UCM, WSU, CMU.
- Recruit LENA users to try out application of analysis tools from the code repository on their own LENA datasets. Incorporate feedback into documentation. UCM.
- Assist independent researchers with accessing HomeBank data and create documentation. UCM, CMU.
- Two conference calls with team of consultants.
- Present at SRCD and IASCL to promote the datasets and code repository.

First Half of Year 3 (Fall and Winter 2017):

- Continue acquiring recordings to be included in the shared datasets. UCM, WSU.
- Continue vetting & stripping private material for public dataset. UCM, WSU.
- Transcribe subsets of the public audio files and add transcriptions to the public dataset. WSU.
- Assist independent researchers with accessing HomeBank data and create documentation. UCM, CMU.
- Continue work documenting and improving interoperability of code in the repository. UCM.
- Submit journal article describing HomeBank and HomeBankCode and illustrating its use to answer a scientific question. UCM, WSU, CMU.
- Two conference calls with team of consultants.
- Submit a proposal to give a workshop at ICIS or arrange a collocated 1-dayworkshop promoting the datasets and code repository.

Second Half of Year 3 (Spring and Summer 2018):

- Secure long-term stewardship board members to create a set of bylaws. UCM, WSU, CMU.
- Wrap up acquisition of recordings to be included in the shared datasets. UCM.
- Wrap up vetting files, strip private material, and add to public dataset. UCM, WSU, CMU.
- Wrap up transcription of subsets of the public audio files and add transcriptions to the public dataset. WSU.
- Wrap up HomeBank and HomeBank Code documentation. UCM, CMU, WSU.
- Revise/resubmit the submitted paper. UCM, WSU, CMU.
- In-person meeting and conference call with consultants and stewardship board members.