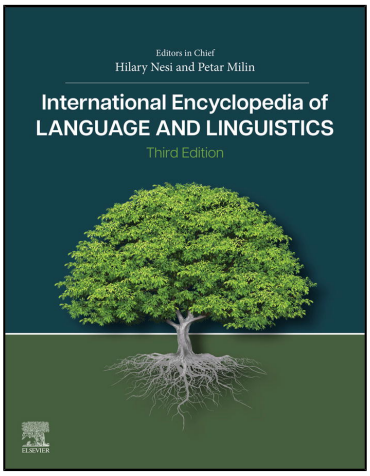

VanDam, M., Warlaumont, A., & MacWhinney, B. (2026). HomeBank: An Online Repository of Daylong Child-Centered Audio Recordings. In: Nesi, H., Milin, P. (Eds.), *International Encyclopedia of Language and Linguistics, 3e*. Vol 12., pp. 490–494. UK: Elsevier. https://dx.doi.org/10.1016/B978-0-323-95504-1.00341-0.
ISBN: 9780323955041

HomeBank: An Online Repository of Daylong Child-Centered Audio Recordings

Mark VanDam[a], Anne Warlaumont[b], and Brian MacWhinney[c], [a] Department of Speech & Hearing Sciences, Elson S. Floyd College of Medicine, Washington State University, Spokane, WA, United States; [b] Department of Communication, University of California, Los Angeles, CA, United States; and [c] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, United States

Key Points

- Child-centered daylong audio recordings provide rich, real-world information about the sounds children produce and hear.
- HomeBank is a repository of thousands of child-centered daylong audio recordings plus their metadata and various annotations.
- Some HomeBank data are fully public and most data are available to researchers who become a HomeBank member.
- HomeBank enables researchers to make use of larger and more diverse real-world child-centered audio datasets, both for scientific discovery and for training machine learning systems.

Abstract

HomeBank (https://homebank.talkbank.org/) is an online repository of long-form child-centered audio recordings available for language, speech, and developmental research. As one of the largest components of the TalkBank system, it contains thousands of hours of recordings collected from real-world (mostly family and home) environments, annotation outputs from diarization software, human listener annotations, demographic data, and associated software for processing and analyzing the large collection of corpora. Recordings and other data are used to study interpersonal and cultural communication, child speech and language development, and other topics in a variety of contexts, and for developing and training automatic audio processing systems.

Introduction

HomeBank is an online collection of thousands of daylong audio recordings in their raw audio formats, the results of automatic speech processing on those audio files, human annotations, metadata associated with the recordings such as demographic details, and software tools to manage the data. The daylong recordings are collected from the perspective of a child in his or her natural home and family environment via a small wireless recorder worn by the child. Researchers using these corpora include those interested in communication, human development, linguistics, automatic analyses, computer science, signal processing, and many other fields.

Body of the Article

Starting in the late 2000s, wearable microelectronics and software for large-scale, long-form speech analysis reached a point where the technology was broadly adopted for use in speech and language development research. Daylong recordings from wearable audio devices were being collected from children and their families in their natural environment. A particularly important organization was the LENA Foundation (Boulder, CO, USA), which developed commercially available audio recording hardware, a small

International Encyclopedia of Language and Linguistics, Third Edition, 2026, 490–494

wearable recorder, and associated software that used automatic speech processing routines to output a time-aligned, diarized description of the audio in terms useful and interesting to the developmental language research community. Typically, recordings are collected in the following manner. After agreeing to the recording, a family is provided with the recording hardware and instructed to begin the recording when the child wakes in the morning. The child wears the recorder in specialized clothes for the duration of the day as they go about their natural daily activities. The parent turns the device off at the end of the day. The recording is then transferred to researchers where it is further processed and managed.

The LENA group's original scientific contribution to the literature described the analyses of 70 h collected from 70 children (Xu et al., 2008a, 2008b). Soon many thousands of recordings were banked in research labs around the world. Researchers varied in their interests (VanDam & De Palma, 2019). Corpora were gathered around many central topics: very young children (Bergelson & Aslin, 2017; Ko et al., 2016; Oller et al., 2010; Walle & Warlaumont, 2015); children with delays and disorders such as hearing loss (Aragon & Yoshinaga-Itano, 2012; VanDam et al., 2012), autism spectrum disorders (Dykstra et al., 2013; VanDam & Yoshinaga-Itano, 2019), and language delays (VanDam et al., 2015); children in bilingual environments (Orena et al., 2020); in cultural context (Ganek et al., 2018); children's interaction with their siblings (Kondaurova et al., 2022); longitudinal development of typically developing children (Gilkerson et al., 2017); preschoolers' math talk exposure (Susperreguy & Davis-Kean, 2016); and many others.

Those research projects depended on large datasets of both raw and processed data, in most cases collected by the interested research teams and stored locally. A notable feature shared by those datasets is that they are inherently composed of rich data that have many variables in complex relationships with rich and varied temporal structure (de Barbaro & Fausey, 2022). Typically, a group collecting a daylong audio dataset intends only to study a small subset of target variables and structures. Given a way to share the data more widely, natural audio recordings can be used as empirical data for a wide range of scientific domains from fields such as psychology, medicine, linguistics, sociology, speech pathology, audiology, second language learning, developmental science, computer science, statistics, data science, signal processing, network analysis, and many more.

These collections of longform audio files are also expensive to collect, difficult to preserve, massive, idiosyncratically structured, and potentially entailed with personal or health information that demands ethical obligations to the participants who contributed those data. HomeBank was conceived and developed by members of a grassroots organization called DARCLE (see below) to capitalize on the strengths of these rich data by facilitating the sharing of these datasets with other researchers—to enable data reuse and larger team science—while attending to the challenges daylong audio present for open data sharing. The extant TalkBank family of corpora, also containing rich spoken language data, were an ideal fit, and HomeBank joined TalkBank as a sub-project in 2015.

Structure of the HomeBank Corpora

HomeBank consists of four principal components. First, there are raw data files, typically. WAV formatted files in each corpus. Recording lengths vary according to the original studies' designs, but typically range from a few hours up to 16 h, which coincides with the memory limits of the LENA recorder or the length of wakeful hours of a child wearing the recorder and their caregiver(s). Results of sound diarization processes are also included. These are typically time-aligned with raw audio files, and most are the processing results of the LENA processing algorithms (resulting in an. ITS, or "interpreted time segments," file). The LENA diarized output yields segment onset and offset times and a one of about 60 apriori labels for each segment identifying, for example, when the target child or another individual (adult female, adult male, another child) is talking, when the TV or radio is on, or periods of silence. Diarization files vary depending on the content from the child and family. The Cougar Corpus, for example, contains 785 daylong recordings, each with a diarization .ITS file. On average, those files have 29,047 labeled segments, ranging from 3939 to 45,192.

Second, each corpus has meta-data associated with the recordings. Metadata depends on the circumstances of the original data collection, often including demographic and personal details. The Lyon Corpus (Le Normand, 2018), for example, contains 49 daylong audio recordings (10–16 h each) from 16 families. An associated metadata file details participant sex, age, birth prematurity, hearing impairment, family history of language or developmental disorders, region of residence, and other details for each recording.

Third, there is a library of software tools for processing the data publicly available on the HomeBankCode GitHub repository (https://github.com/homebankcode/). The tools were created by users with a wide variety of needs for tools not already available elsewhere. For example, there are software solutions for parsing diarization files, extracting acoustic phonetic features, and splicing/combining portions of raw audio files. Software in the repository include scripts in MATLAB, Python, R, C, and other languages.

Fourth, there is a community of interested contributors, Daylong Audio Recordings of Children's Linguistic Environments (DARCLE; https://darcle.org/). This community was established to support the complex research enterprise of working with daylong child-centered audio data and was the community from which HomeBank originated. DARCLE is a worldwide research consortium with free and open access to all its resources. The community meets virtually on a regular basis to discuss issues of interest to the larger community of researchers. DARCLE also includes resources for New Investigators and students to engage with the content and other interested researchers.

Classification of HomeBank Corpora

HomeBank has three types of corpora: public, member, and limited. Public corpora are open to anyone on the internet to download, access, or use without restriction. Public corpora have been curated to ensure that the participants who have given their data are treated in an ethical way, conforming to the researchers' original IRB or other ethics board protocols and participants' data sharing preferences. The raw audio and associated transcripts or diarization files are examined so personal details or identifying characteristics are not present (VanDam et al., 2016). For example, The Fausey Trio Public Corpus (Fausey & Mendoza, 2018b), itself a subset of the Fausey Trio Corpus (Fausey & Mendoza, 2018a), contains three daylong recordings from one participant and does not include full demographic or diarization details.

Member corpora are available to HomeBank members (see below) only. Members are required to agree to ethical treatment of participant data, but otherwise have unrestricted access to this category of corpora. By number of recordings and total hours of recordings, this is the largest category of data in HomeBank. Because of the total volume of recordings, it is often impossible to fully vet recordings in their natural context. For example, the San Joaquin Valley Corpus (Warlaumont et al., 2024) contains 166 daylong audio recordings from 56 participants. Due to the large number and length of recordings, only 11 of the audio recordings have thus far undergone the full double vetting and scrubbing required for public sharing per the contributors' IRB protocol; the full corpus has not been reviewed for content but is available for use by all HomeBank members.

Limited corpora are available only by individual agreement between the researcher responsible for the data and collaborators seeking to use the data. In some cases, use may require certain Institutional Review Board considerations or other arrangements. These data are treated specially in accordance with original data collection procedures often due to the sensitive nature of the participants who contributed their data. For example, the Ellis Corpus (Ellis, 2024) consists of kindergarten teachers in southeast Michigan during the school day. Due to privacy concerns, the data are in the restricted Limited category of HomeBank corpora.

Membership

HomeBank membership is open to any researcher. Membership is initiated by the prospective member who provides basic contact and professional information and evidence of training on the ethical and responsible treatment of human participant data. HomeBank conducts a live interview with members to describe the resource and introduce the member to the project. Membership is granted to the lead researcher or principal investigator and collaborators, students, and staff are managed by the Member (with notification and verifying ethics training paperwork updated with HomeBank central records). Members have unique login credentials to access the corpora and online resources. The majority of members are academic researchers and their students. Members represent worldwide academic and industry interests.

TalkBank and HomeBank Web Interface

HomeBank benefits from being part of the larger TalkBank family of resources (MacWhinney, 2014) and from inheriting TalkBank structural and formatting conventions. In addition to seamless integration with CLAN and PHON software, HomeBank makes sitewide use of TalkBankDB and TalkBank Browser. TalkBankDB is a database tool that allows users to sort, filter, and organize the data quickly by useful variables of interest. With the web interface, users select for participants, age, file types, details of transcripts, and many other variables. Search results can be downloaded in. csv files and downloads and searches can be performed from within R or Python using the provided APIs (application programming interfaces). TalkBank Browser is a user-friendly, online interface to interact with all of the HomeBank data, including media files and transcripts. In this interface, the user can select an audio file and associated transcript (such as hand-coded transcript, automated diarization file, or other) and interact with the file, including listening to audio playback, within the browser. In addition, CLAN commands can be run directly on the data using commands in the TalkBank Browser. Finally, Collaborative Commentary, a system for adding annotations to transcripts and validating coding schemes, can be used with HomeBank data, including data with transcripts created through ASR.

HomeBank's Impact and Future

HomeBank is an active scientific resource used by hundreds of members with a substantial presence in the empirical scientific child development literature (e.g., Bergelson et al., 2023; Cychosz et al., 2021; Ellwood-Lowe et al., 2021). HomeBank data have also been used as training and validation data for the purposes of developing open-source machine-learning-based tools for identifying events of interest within daylong child-centered audio recordings (e.g., Räsänen et al., 2021; Schuller et al., 2017).

Several initiatives are underway to capitalize further on the very large database of rich data. For example, Secure HomeBank is a tool in development intended to allow researchers to initiate analyses (such as acoustic phonetic research) remotely without directly accessing the underlying audio. This will respect the agreements the original researchers have made with participants while still promoting scientific investigation. Another goal is to expand and refine the end-user interface of the resource to reduce the technical burden of access. Additionally, there is the enduring goal of gaining additional rich data corpora to expand the scientific

usefulness of HomeBank. There is a TalkBank Board of Governors to moderate and improve the overall resources of TalkBank as well as a HomeBank Stewardship Committee dedicated to maintaining the integrity and usefulness of HomeBank specifically. These groups meet regularly to improve and promote the resources and scientific inquiry.

Conclusion

HomeBank provides rich resources for the study of a variety of important aspects of language development and socialization. The system has benefited from multiple advances in recording technology, data-sharing, and corpus analysis tools. Ongoing advances in each of these areas will contribute further to the richness, coverage, usability, and importance of HomeBank.

Acknowledgments

This work was supported by National Science Foundation grants (1539133, 1539010, 1529127 and 1539129/1827744), National Institutes of Health NIDCD DC009560-01S1, the Washington Research Foundation, and a James S McDonnell Foundation Scholar Award to Anne S. Warlaumont. We would also like to thank the HomeBank Stewardship Committee as well as all the HomeBank data and code contributors.

References

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*(04), 340–353. https://doi.org/10.1055/s-0032-1326918

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences, 114*(49), 12916–12921. https://doi.org/10.1073/pnas.171296611

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramírez-Esparza, N., Hamrick, L., Marklund, E., Kalashnikova, M., Guez, A., Casillas, M., Benetti, L., van Alphen, P., & Cristia, A. (2023). Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences, 120*(3), Article e2300671120. https://doi.org/10.1073/pnas.2300671120

Cychosz, M., Cristia, A., Bergelson, E., Casillas, M., Baudet, G., Warlaumont, A. S., Scaff, C., Yankowitz, L., & Seidl, A. (2021). Vocal development in a large-scale crosslinguistic corpus. *Developmental Science, 24*(5), Article e13090.

de Barbaro, K., & Fausey, C. M. (2022). Ten lessons about infants' everyday experiences. *Current Directions in Psychological Science, 31*(1), 28–33.

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. https://doi.org/10.1177/1362361312446206

Ellis, A. (2024). *Homebank English Ellis Corpus*. https://doi.org/10.21415/CXHA-CM62

Ellwood-Lowe, M. E., Foushee, R., & Srinivasan, M. (2021). What causes the word gap? Financial concerns may systematically suppress child-directed speech. *Developmental Science, 25*(1), Article e13151. https://doi.org/10.1111/desc.13151

Fausey, C. M., & Mendoza, J. K. (2018a). *FauseyTrio HomeBank Corpus*. https://doi.org/10.21415/T5JM4R

Fausey, C. M., & Mendoza, J. K. (2018b). *FauseyTrio-Public HomeBank Corpus*. https://doi.org/10.21415/T56D7Q

Ganek, H., Smyth, R., Nixon, S., & Eriks-Brophy, A. (2018). Using the Language ENvironment Analysis (LENA) system to investigate cultural differences in conversational turn count. *Journal of Speech, Language, and Hearing Research, 61*(9), 2246–2258. https://doi.org/10.1044/2018_JSLHR-L-17-0370

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., … Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology, 26*(2), 248–265. https://doi.org/10.1044/2016_AJSLP-15-0169

Ko, E.-S., Seidl, A., Cristia, A., Reimchen, M., & Soderstrom, M. (2016). Entrainment of prosody in the interaction of mothers with their young children. *Journal of Child Language, 43*, 284–309.

Kondaurova, M. V., Zheng, Q., VanDam, M., & Kinney, K. (2022). Vocal turn-taking in families with children with and without hearing loss. *Ear and Hearing, 43*(3), 883–898. https://doi.org/10.1097/aud.0000000000001135

LeNormand, M. T. (2018). *Lyon HomeBank Corpus*, 21415/T58P6Q.

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. New York: Psychology Press. https://doi.org/10.4324/9781315805672

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, 107*(30), 13354–13359. https://doi.org/10.1073/pnas.1003882107

Orena, A. J., Byers-Heinlein, K., & Polka, L. (2020). What do bilingual infants actually hear? Evaluating measures of language input to bilingual-learning 10-month-olds. *Developmental Science, 23*(2), Article e12901. https://doi.org/10.1111/desc.12901

Räsänen, O., Seshadri, S., Lavechin, M., Cristia, A., & Casillas, M. (2021). ALICE: An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. *Behavior Research Methods, 53*, 818–835. https://doi.org/10.3758/s13428-020-01460-x

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., Warlaumont, A. S., Hidalgo, G., Schnieder, S., Heiser, C., Hohenhorst, W., Herzog, M., Schmitt, M., Qian, K., Zhang, Y., Trigeorgis, G., Tzirakis, P., & Zafeiriou, S. (2017). The INTERSPEECH 2017 Computational paralinguistics challenge: Addressee, cold, & snoring. *Proceedings INTERSPEECH*, 3442–3446. https://doi.org/10.21437/Interspeech.2017-43

Susperreguy, M. I., & Davis-Kean, P. E. (2016). Maternal math talk in the home and math skills in preschool children. *Early Education and Development, 27*(6), 841–857. https://doi.org/10.1080/10409289.2016.1148480

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. https://doi.org/10.1093/deafed/ens025

VanDam, M., & De Palma, P. (2019). A modular, extensible approach to massive ecologically valid behavioral data. *Behavior Research Methods, 51*, 1754–1765. https://doi.org/10.3758/s13428-018-1167-8

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, 36*(4), e146–e152. https://doi.org/10.1097/AUD.0000000000000138

VanDam, M., Warlaumont, A., MacWhinney, B., Soderstrom, M., & Bergelson, E. (2016). *Vetting manual: Preparation of recordings for unrestricted publication in HomeBank*. https://doi.org/10.21415/T56H4M

VanDam, M., & Yoshinaga-Itano, C. (2019). Use of the LENA autism screen with children who are deaf or hard of hearing. *Medicina, 55*(8), 495. https://doi.org/10.3390/medicina55080495

Walle, E., & Warlaumont, A. S. (2015). Infant locomotion, the language environment, and language development: A home observation study. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2577–2582). Santa Barbara, CA: Cognitive Science Society.

Warlaumont, A. S., Pretzer, G. M., Mendoza, S., Schneider, S., Mutrie, J., Lopez, E. A., & Kello, C. T. (2024). *San Joaquin Valley HomeBank Corpus*. https://doi.org/10.21415/T54S3C

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

Xu, D., Yapanel, U., Gray, S., Gilkerson, J., Richards, J., & Hansen, J. (2008b). Signal processing for early child speech language development. In *Paper presented at the* First Workshop on Child, Computer, and Interaction (WOCCI). *Chania, Crete, Greece*.

Relevant Websites

TalkBank. https://www.talkbank.org/
HomeBank. https://homebank.talkbank.org/
DARCLE. https://darcle.org/
Github code repository. https://github.com/homebankcode/