Daylong data: Raw audio to transcript via automated & manual open-science tools

John Bunce (john.bunce@umanitoba.ca)
Department of Psychology, University of Manitoba
190 Dysart Road, Winnipeg, MB R3T 2N2 Canada

Elika Bergelson (elika.bergelson@duke.edu)
Department of Psychology and Neuroscience, Duke University
417 Chapel Drive, Campus Box 90086, Durham, NC 27708-0086

Anne Warlaumont (warlaumont@ucla.edu)
Department of Communication, University of California, Los Angeles
2225 Rolfe Hall, UCLA, Box 951538, Los Angeles, CA 90095

Marisa Casillas (marisa.casillas@mpi.nl)
Language Development Department, MPI for Psycholinguistics
Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands

Abstract

Several of the central questions in language, social cognition, and developmental research focus on the roles of input, output, and interaction on learning and communication. While it has become easy to collect long-form recordings, getting useful data out of them is a more daunting task. Across four mini-sessions, this tutorial aims to address pre- and post-data collection concerns, and provide a hands-on introduction to manual and automated annotation techniques. Attendees will leave this tutorial with resources and concrete experience for collecting, annotating, and sharing/archiving naturalistic recordings, including specific open-science practices relevant for these data.
Keywords: daylong recordings; natural language; speech technology; automated annotation; open science

Introduction

The ability to record and efficiently analyze everyday talk from a variety of different populations is crucial for many topics in language science, including: variation in children's linguistic input, distributional patterns of language in adult speech, atypical speech patterns for medical diagnosis, and more (e.g., see Casillas and Cristia (under review) for a review). However, even the most basic facts about everyday speech experience have remained elusive given the technological constraints of capturing and analyzing daylong speech for large samples of participants. In the last two decades, the LENA™ system has emerged as a potential solution to this methodological gap (see Ganek and Eriks-Brophy (2018) for a review). However, due to its costs and proprietary, aging technology, LENA™'s usefulness is increasingly limited.

In this half-day tutorial we will describe a new approach for getting the most from daylong recordings; one that uses community-based norms to support researchers at every step, from ethics review and initial data collection to automated analysis, manual annotation, and data archival. The tools and databases we include are all open-source and oriented toward usability on new populations and new technical challenges—an ideal next step to enable researchers to tackle new scientific questions about everyday language use. These tools has developed out of the ACLEW project (`http:\sites.google.com/view/aclewdid/home`).

Tutorial aims

This tutorial is focused on facilitating current research using daylong recordings while also boosting the future development of *even better* tools for the collection, annotation, and analysis of daylong recordings.

Our first aim is to **lower the barrier to using daylong recordings for language research**. Many researchers who are interested in this method are held back from doing so because there is no clear cost- and time-efficient way to annotate the data. We hope to allay some of these concerns by introducing a set of tools and techniques participants can use to extract usable data from their recordings. We will provide a hands-on training session demonstrating how to use our ACLEW audio-processing pipeline (automated tools for exploring voice activity, utterance segmentation, speaker diarization, and speech rate estimation) and manual annotation framework suitable for cross-corpus comparison. All software is free, open-source, and multi-platform.

Our second aim is to **promote an open-science framework for natural language data**, with an eye toward improving access to shared data and comparative analysis. The daylong recording community is just getting off the ground (HomeBank; VanDam et al., 2016), and there is vast potential for scientific advancement if more researchers were to participate. To demonstrate the benefits of data sharing and re-use for daylong recordings, we will show how the use of unified tools and annotation templates can lead to new breakthroughs in comparing natural language environments across cultures. Our motivation is that the long-term non-commercial success of our toolkit depends on an active community of users. Active users contribute new training data, give feedback on quality, and make requests for new functionality. We therefore hope to convince researchers that these tools can meet their immediate analytic needs while also persuading them to invest in the community so that we can establish the mega corpora necessary for continued tool improvement.

Participants

This tutorial is intended for researchers at all levels of experience who are interested in the collection, analysis, curation, and computational modeling of natural language data. While the tutorial will be accessible to a general CogSci audience, we also hope to attract participants who are interested in daylong recordings but daunted by the prospect of collecting or processing them. We also encourage participation by researchers who have already invested in daylong recordings and are looking for new ways to utilize them. Indeed, as part of DARCLE we have a commitment and track record of supporting new investigators (`http://darcle.org/newInvestigators.html`).

Learning outcomes

After this tutorial, participants will be able to (1) assess the pros and cons of using naturalistic recordings for their research questions, (2) locate, use, and adapt our online, self-guided tutorials and templates for creating machine-friendly annotations, (3) download, install, run, and interpret the output provided by the (open source) audio-processing software, and (4) understand how to gain access to and use HomeBank, a repository for daylong audio recordings.

Tutorial structure

This half-day hands-on tutorial will introduce: issues surrounding daylong recording collection, a standardized manual annotation process, the use of automated annotation tools, and best practices for data archiving. This will be organized into four sessions (separated by 5-min breaks). Participants will work with sample media file to get hands-on experience in each session.

Session 1. Pre-data collection concerns (25 min) A brief introduction to the method, its costs and benefits, and what to consider before collecting data. Topics include: how to decide whether daylong recordings are suitable for the research question, considerations when applying for ethical approval, and off-the-shelf hardware and software options. We will relate these topics to individual research interests.

Session 2. Manual annotation (55 min) A 3-part interactive training session introducing participants to manual annotation in the machine-friendly template we have developed for ELAN (Casillas et al., 2017). Part 1 focuses on the basic setup of the annotation scheme. Part 2 focuses on the use and adaptability of the annotation conventions. Part 3 focuses on the annotator training standards and reliability estimation using the automated tools provided by ACLEW.

Session 3. Automated annotation (55 min) An interactive tour of the ACLEW automated tools package. Each tool will be introduced and demonstrated with example media files. We will also take this opportunity to demonstrate the value of adding new training and testing data and will open the floor to discussion about future tool development.

Session 4. Archiving and community (25 min) A brief discussion focused on the issues surrounding the long-term storage of daylong recordings. We will also discuss efficient and accessible ways to share data, annotations, and analysis, and review the benefits of open-science practices.

Learning materials

Participants will need an Internet-connected laptop and a pair of headphones. The organizers will create an OSF page with links to all training materials and instructions for future use. Although sample data will be provided, participants are encouraged to bring their own data to demonstrate the challenges of different research questions using daylong audio.

Tutor credentials

The materials and instruction for this tutorial will come from the cognitive scientists and software developers who created the tools being covered. Collectively, they have expertise in training dozens of researchers (undergraduate to PhD) on the steps covered in sessions 1–4. That said, this tutorial will be the very first to cover the end-to-end use of this pipeline for researchers working on daylong audio recordings.

Summary of significance

The study of everyday talk is fundamental for understanding the relationship between cognition, culture, and language. Recent technological advancements afford researchers the ability to study everyday language on a much larger scale than before, but these technologies are challenging and therefore remain somewhat underutilized. We aim to further the use and usefulness of this technology by spreading knowledge of how to effectively employ it and by facilitating the continued improvement of the associated tools for language science.

Acknowledgments

This work was supported by a TransAtlantic Platform "Digging into Data" collaboration grant (ACLEW: Analyzing Child Language Experiences Around the World) and an NWO Veni Innovational Research Scheme (275-89-033) to MC.

References

Casillas, M., Bunce, J., Soderstrom, M., Rosemberg, C., Migdalek, M., Alam, F., ... Garrison, H. (2017). *Introduction: The ACLEW DAS template [training materials].* Retrieved from `https://osf.io/aknjv/`

Casillas, M., & Cristia, A. (under review). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *XX*, *XX*, XX–XX.

Ganek, H., & Eriks-Brophy, A. (2018). Language ENvironment analysis (LENA) system investigation of day long recordings in children: A literature review. *Journal of Communication Disorders*, *72*, 77–85.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., De Palma, P., & MacWhinney, B. (2016). *HomeBank: An online repository of daylong child-centered audio recordings.*