UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Casillas, M., and Cristia, A. (2019). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra: Psychology,* 5(1): 24. DOI: https://doi.org/10.1525/collabra.209

REVIEW ARTICLE

A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings

Marisa Casillas* and Alejandrina Cristia†

Recent years have seen rapid technological development of devices that can record communicative behavior as participants go about daily life. This paper is intended as an end-to-end methodological guidebook for potential users of these technologies, including researchers who want to study children's or adults' communicative behavior in everyday contexts. We explain how long-format speech environment (LFSE) recordings provide a unique view on language use and how they can be used to complement other measures at the individual and group level. We aim to help potential users of these technologies make informed decisions regarding research design, hardware, software, and archiving. We also provide information regarding ethics and implementation, issues that are difficult to navigate for those new to this technology, and on which little or no resources are available. This guidebook offers a concise summary of information for new users and points to sources of more detailed information for more advanced users. Links to discussion groups and community-augmented databases are also provided to help readers stay up-to-date on the latest developments.

Keywords: speech environment; natural language; daylong recordings; annotation; LENA

1. Introduction

Daily practices of communication, such as sharing news, making requests, and answering questions, shape the processes by which we use language to communicate with others. These everyday patterns of language use, such as how often words are used in interaction (and who uses them), influence how we produce and comprehend language. By that token, a priority for those investigating language and cognition should be to track the ways in which people typically encounter language. Accurate documentation of natural language environments would enable us to more effectively hypothesize mechanisms for language learning and processing, and to test whether our theories about language scale up to everyday use. Until recently, we did not have access to technology that allowed us to reliably and efficiently collect and analyze data on people's daily language experiences. But these days the hardware for audio- and video-recording is increasingly cheap, small, and wearable, and is often combined with other types of simultaneously recorded data (e.g., GPS, heart rate, body acceleration). Software for audio manipulation, annotation, and analysis is also improving rapidly, helping researchers to find meaningful moments in large, complex datasets without first requiring them to spend months manually annotating their data. By bringing these technologies together, we have already begun to get a glimpse into the language used in peoples' daily lives, particularly with respect to children's language environments (see Section 1.1). With the refinement of these technologies, we can expect to see major advances in models of language processing and learning.

What can we do to usher this progress along? Today's data collections become the basis for tomorrow's theories and tools. As researchers who work on language, our current task is to carefully consider how we can collect, annotate, store, and analyze data such that we can optimally produce informative work in the present while also benefiting our future research.

In the rest of this Introduction we will briefly review past research using LFSE recordings with child and adult populations, discuss what is needed in future LFSE research, and then give an outline for the remainder of the paper.

1.1. Child language research

LFSE recordings are becoming a central method for studying how children learn language. Much research on language development has focused on what children hear and say because children's early language experience is taken to be the "input" they use to infer the grammatical structure and lexicon of their language(s) (e.g., Frank, Braginsky, Marchman, & Yurovsky, 2019; Cartmill et al., 2013; Huttenlocher, Waterfall, Vasilyeva, Vevea, & Hedges,

* Max Planck Institute for Psycholinguistics, Nijmegen, NL

† Laboratoire de Sciences Cognitives et Psycholinguistique, Dept d'Etudes Cognitives, ENS, PSL University, EHESS, CNRS, Paris, FR

Corresponding author: Marisa Casillas (marisa.casillas@mpi.nl)

2010; Hoff, 2006). Major research questions for child language development include: Who talks to the child, how often, and what do they say? What influence does the child's own speech have on what they hear? How do those measures change with age? Could these behaviors be used to effectively detect language delays/disorders and the effects of interventions? In order to propose and test mechanisms for language learning, researchers need to know what information is available to children in their language environments.

Traditionally, developmental language researchers have drawn their assumptions about the speech children hear from short recordings of naturalistic interaction, often made under semi-controlled conditions, such as interaction in a free-play area in the lab, or in children's homes. These sampling methods give researchers an informationally rich, but otherwise narrow view into the total range of communicative situations children observe and participate in from waking to sleeping. With the introduction of the LENA recording device (https://www.lena.org/; Greenwood, Thiemann-Bourque, Walker, Buzhardt, & Gilkerson, 2010) in the late 2000's, researchers were finally able to both continuously record children's whole waking days with a wearable microphone, thereby putting the child's own perspective in the spotlight. LENA's recording device also comes with software that automatically detects and classifies stretches of speech into noise-type categories (e.g., target child speech, female adult speech, etc., plus several non-speech categories; see Ganek & Eriks-Brophy 2018a, for an overview of the system). LENA-based research has been conducted to different extents in a range of communities in North America, Europe, and East Asia; (Bergelson & Aslin, 2017; van Alphen, Meester, & Dirks, 2017; Busch, Sangen, Vanpoucke, & Wieringen et al., 2017; Schwarz, Botros, Lord, & Marcusson, 2017; Canault, Normand, Foudil, Loundon, & Thai-Van, 2016; Ganek & Eriks-Brophy 2018a; 2018b; Elo, 2016; Gilkerson et al., 2015; Soderstrom & Wittebolle, 2013; VanDam, Ambrose, & Moeller, 2012; Weisleder & Fernald, 2013).

By combining the LENA system's automated output with manual coding—for example, by adding information about who the speaker is talking to—researchers have been able to make several basic characterizations of children's daily speech environments. The findings based on LENA data so far suggest that there is enormous variability within and across the groups whose language environments have been studied closely (e.g., Bergelson & Aslin, 2017; Ganek & Eriks-Brophy, 2018b; Soderstrom & Wittebolle, 2013; VanDam et al., 2012; Weisleder & Fernald, 2013; see Ganek & Eriks-Brophy, 2018a for a review). LENA-based studies have also shown that aspects of both the quantity and quality of the speech children hear relates to linguistic development (e.g., Weisleder & Fernald, 2013), and that speech directly addressed to the child might have greater influence on their development than the total amount of speech they hear, especially when that directed speech is designed to engage the child (e.g., Ramírez-Esparza, García-Sierra, & Kuhl, 2014; 2017; Romeo et al., 2018). Though LENA's system was originally designed for use with American English, it has been recognized by researchers as a useful tool in several other language communities. In these cases, researchers have worked to validate its use for their languages and circumstances of interest (e.g., Marchman, Martinez, Hurtado, Grüter, & Fernald, 2016; Canault et al., 2016; Woynaroski et al., 2016; Gilkerson et al., 2015).

Overall, LFSE recordings have been used very actively in the last decade to study child language. Their use to study adult language has grown somewhat more slowly, as we will see below.

1.2. Adult language research

We still know surprisingly little about language use in the daily life of adult speakers. Prior research, sometimes relying on data from mobile phones and social media, has focused on using typical language patterns to investigate language-based measures of interactional patterns (Lee et al., 2013), personality (Park et al., 2015), mental health (e.g., Karam et al., 2014), and also to design technologies such as voice command software for individuals with speech and motor disabilities (Nicolao, Christiansen, Cunningham, Green, & Hain, 2016). However, LFSE recordings have important insights to add to these lines of applied research—there are still many unanswered basic questions about adults' language experience: What proportion of minutes per day are spent in face-to-face conversation? How many speakers are typically encountered in a day? What mediates the variation in those measures across different lifestyles and communities, and at different points in the lifespan? Could the quantity and quality of linguistic interactions be used as predictors for neurocognitive status, including measuring the effects of interventions? The quantity and diversity of language encountered over the course of the day shapes our cognitive processes for producing and comprehending language, but we do not yet know precisely what is affected, and how.

A few pioneering groups (e.g., Mehl, Ramírez-Esparza, Hansen, and each of their colleagues) have begun to investigate the daily language patterns of adults using LFSE technology. For example, Mehl, Pennebaker, and colleagues have been using a device called the EAR (Electronically Activated Recorder; Mehl, Pennebaker, Crow, Dabbs, & Price, 2001; Mehl, 2017; see also Rodríguez-Arauz, Ramírez-Esparza, García-Sierra, Ikizer, & Fernández-Gómez, 2018) to get a better grip on adults' daily linguistic environments in a few different settings. The EAR is a wearable system (now a smartphone app, "iEAR") that does not sample continuously; instead, it records 30- or 50-second snippets of audio every 12 minutes (see more details in Mehl et al., 2001; Mehl, 2017). The snippets must then be manually transcribed and annotated before they can be analyzed or related to other behavioral measures. These other behavioral measures have included, for example, self-reports on mental health, personality, and social measures (e.g., Rodríguez-Arauz et al., 2018; Robbins, Mehl, Holleran, and Kasle, 2011; Mehl, Vazire, Holleran, & Clark, 2010; Ramírez-Esparza, Mehl, Álvarez-Bermúdez, & Pennebaker, 2009). Using this method, Mehl and colleagues (2007; 2010) have been able to make initial estimates of some basic language measures, including

average variance in the amount of speech produced during a day by participants (Mehl, Vazire, Ramírez-Esparza, Slatcher, & Pennebaker, 2007) and the relative frequency of "small talk" vs. "substantive talk" (Mehl et al., 2010). Notably, a similar snippet-based method can be used with the LENA device (see, e.g., Ramírez-Esparza et al., 2014; 2017), which has the benefit of continuous recordings from which talkative snippets can be selected and transcribed—an efficient approach for those focused on verbal or other interactional behavior characteristics. Ultimately, researchers using LFSE recordings with adults have so far typically been interested in relating language measures to non-linguistic aspects of participants' social lives (e.g., happiness, overall sociality, etc.), and so there is still much to learn about the fundamental characteristics of and variance in adult language environments.

This previous work using LFSE recordings with both adult and child participants has laid a solid foundation for future work on language use in natural contexts. However, as LFSE recording technologies evolve, researchers who are interested in species-wide language processes (i.e., human linguistic cognition, not limited to a single population) will also need to carefully consider how to most effectively capture an ecologically valid sample of the world's language communities.

1.3. Goals for future LFSE research

Future work with LFSE recordings faces several challenges, as follows. In order to gain a comprehensive sample of human language use (i.e., good estimates of variance in adult and child language environments) we must both increase the diversity of represented human linguistic communities with LFSE data while also decreasing our reliance on manual annotation, which is too slow and sometimes altogether impossible to work at large scales (i.e. thousands or, perhaps eventually, millions of hours of data).

1.3.1. Increasing diversity

Daily linguistic experiences are bound to be affected by the wide-ranging cultural, economic, and linguistic differences that exist between speakers living in different communities around the globe. Yet most of the participant samples used in LFSE research so far have primarily come from healthy adults and typically developing children in industrialized (often Western) settings. We need data from a more diverse sample of human language environments and clinical conditions to discover human-wide principles underlying language development and language processing.

This sampling bias holds for psychology at large (Henrich, Heine, & Norenzayan, 2010), but also specifically for the sub-fields of adult psycholinguistics (Norcliffe, Harris, & Jaeger, 2015), developmental psychology (Nielsen, Haun, Kärtner, & Legare, 2017), and developmental language studies (e.g., Slobin, 2014). For example, while some collections of naturalistic language data have attracted linguistically diverse recordings (e.g., CHILDES; https://childes.talkbank.org/; MacWhinney, 2000), populations from industrialized, often Western populations are still overrepresented relative to the global set of cultural communities existing today (see also the DOBES Archive). Among studies that use LFSE recordings of everyday language, there have been a few comparative samples of cultural groups from industrialized societies (e.g., Ramírez-Esparza et al., 2009; Weisleder & Fernald, 2013; the ongoing ACLEW project, ACLEW Project, 2018), with a handful more from traditional indigenous and/or non-industrial societies (Casillas, Brown, & Levinson, under review; Scaff, Stieglitz, Casillas, & Cristia, in preparation; Abels & Vogt, 2018). However, there is still a world of cultural and linguistic diversity in language experience that has yet to be captured with this method.

Comparative cross-linguistic and cross-cultural studies of natural language use must be approached with caution because, in many cases, linguistic and cultural phenomena cannot be directly compared. A primary challenge moving forward will then be to further develop computational tools that allow us to reliably extract measures (e.g., minutes of nearby speech) from recordings in as unbiased a manner as possible. Another approach one could use to validly compare LFSE recordings directly would be to work with similar (ideally culturally and linguistically related) communities that differ in some crucial way, for example, in their extent of contact with the post-industrial Western world (see Pye, 2017).

1.3.2. Decreasing manual annotation

Many of the studies mentioned above rely on manual annotation, in part or in whole, to create meaningful input for their analyses. A continued reliance on manual annotation is impractical on the scale of LFSE recordings, which easily accumulate to thousands of hours of data. Even well-funded projects with unrestricted access to human annotators would benefit in time savings and greater comparability of output by using automated tools as part of their data processing pipeline (e.g., finding segments of speech vs. silence).

Basic automated tools for everyday speech corpora are still in their infancy. While LENA's software has been able to produce automated annotations since its inception in the late 2000's, its software is proprietary, outdated, and the output is limited to the information that the LENA organization originally decided to focus on (e.g., broad speaker type like "female adult" rather than individual speakers). Alternative tools for automated annotation are an active area of research. For example, Hansen and colleagues (e.g., Ziaei, Sangwan, Kaushik, & Hansen, 2015) have been trying to further develop tools for data collected with a LENA recording device on an wide range of tasks, including: speech activity detection (finding stretches of speech vs. non-speech), diarization (assigning speech stretches to speaker sources), finding specific words in the audio, word-counting, and identifying the speaker's location/activity. However, the recordings they are working on are limited to a single participant's range of daily activities, and their tools are, for the most part, not accessible to other researchers.

While the first few automated processing steps (such as finding where someone is talking versus silent) may seem trivial, research on this topic shows that modern systems do not perform well on LFSE recordings. In fact, the performance of commercially available speech technology

systems, which appear to work much better, is possible because those technologies are tuned to specific settings. For instance, digital assistants (e.g., Siri or Google) are very effective at recognizing and responding to a user's spoken request about nearby restaurants, because: (a) they can access a great deal of contextual information including the user's speech patterns, GPS location, time of day, etc., (b) they have access to a massive quantity of user data, allowing them to pool feedback (e.g., user tries the search again, which means the previous results were not good), and (c) their specialized engineers make improvements to their software task by task. Notably, the tasks that these engineers prioritize (e.g., retrieving restaurant information) are only sometimes in-line with the goals of researchers studying human language use for its own sake. LFSE recordings pose unique challenges because there is often quite a lot of background noise and multiple talkers, all of which are recorded with a single microphone on a single person's body. This added variability from more types of talkers in more types of language environments compounds the difficulty of accurate automated annotations.

The ideal scenario is for researchers who are working with LFSE recordings to also be involved in the development of those associated speech technologies—in that way researchers can help ensure that LFSE annotation systems are robust to the variability of real-life language experience. There is a lot of room for current technologies to improve. Though many researchers may hope for automated transcription, more likely goals for the short term are much simpler, such as tracking time talking to family members. Researchers collecting data can encourage faster development of these and other tools by using them on their data when possible, and by ensuring that the manual annotations they do generate can be used as training materials for future tool development, e.g., via community-maintained annotation formats like CHAT (MacWhinney, 2000) and the DAS (Casillas, Bergelson, et al., 2017), used with ELAN (Sloetjes & Wittenburg, 2008). We return below to the availability and performance of such computational tools for the kinds of data that interest our readers.

1.4. Our aim

As we will see, researchers planning to collect or analyze LFSE recordings face a number of unique challenges. The first pertains to the ethical and legal aspects associated with data that cannot be de-identified, such as voice recordings. This is an issue that affects all stages of the research, from its conception and submission to an ethics committee, to the archival of the data. A second set of challenges relates to the fast-changing, ever-improving, and yet always seemingly suboptimal technology required to collect and analyze LFSE datasets. A third set of challenges concerns the design and validation of measurements extracted from these datasets. In this paper, we aim to help potential users of LFSE recording methods to decide whether such data can feasibly provide answers to their research questions. For those readers who decide that LFSE recordings are suitable for their research, we provide a guide to help make the most of this promising and challenging set of technologies.

Given the authors' fields of expertise, this manuscript may be most relevant for psychologists, linguists, and anthropologists interested in understanding the mechanisms underlying first language development and early social development. We hope that this guide will also be informative for researchers working in other domains, including auditory ecology and non-industrial (or other understudied) language contexts; in particular, we can foresee exciting extensions of this technology for field linguists who could use this technology for language documentation and/or sociolinguistic investigation throughout the lifespan. Similar approaches are also gaining popularity in the medical sciences, as clinical psychologists, medical doctors, language pathologists, social workers, and other practitioners consider using LFSE recordings to measure social and/or verbal behavior in everyday life of patients, the elderly, and other populations (see, e.g., Mehl et al., 2010). Notably, however, the focus for the rest of our paper will be on fundamental research using LFSE technology and not its application by these professional groups.

2. When to use LFSE recordings

In this section we encourage researchers to begin by thinking about whether LFSE recording methods are right for their research question. Although LFSE recordings have benefits (e.g., larger samples, naturalistic recording contexts, some reliable automated annotations), they can also be intrusive for participants and typically require an enormous investment of time and resources to answer even simple questions. Those considering an LFSE recording-based project should first carefully consider how they will satisfy the basic requirements for collecting, analyzing, and storing LFSE recordings, and also critically examine whether their research question could be answered through more efficient means.

While collecting LFSE recordings is very easy today, analyzing them remains challenging. Even apparently simple tasks, like deciding whether a stretch of audio is human speech or engine noise, still show accuracy scores substantially below 100%. At this time, automated assignment of speech to individual speakers and automated transcription of LFSE recordings cannot feasibly replace manual annotation—not even for the populations with the most available training data for the algorithms to learn from (American English; see Section 4 for more details). Nonetheless, we believe that there are other automated annotation tasks, which are both technically possible and scientifically informative, that make the LFSE recording method worthwhile. For example, with LFSE recordings one can (a) extract ecologically valid (but short) samples for analysis at broader or finer levels of detail; or (b) run automatic analysis routines to provide broad numerical descriptions of language environments, such as overall quantity of speech produced by or near the person wearing the recorder.

A key question for researchers to ask themselves is whether LFSE recordings will be more useful compared to alternatives (such as short, targeted recordings, standardized tests, etc.). LFSE recordings are most useful for studying broad linguistic or behavioral phenomena that are highly

frequent and have salient acoustic signatures. To take a few specific examples, characteristics of the overall acoustic profile of the recording, the quantity of high-pitch noise, and the quantity of speech produced by the person wearing the recording device are all measures that have a salient acoustic signature and that occur with high frequency throughout the recording. As a consequence, all three of these measures can be extracted with reasonable accuracy.

In contrast, if our research question focused on a rare phenomenon (e.g., one that occurs only in specific conditions), or one that has no salient acoustic signature, then, from the machine's perspective, the phenomenon is a tiny needle in an immense LFSE recording haystack and it will be difficult to accurately identify automatically. For instance, imagine that one is interested in the use of a certain linguistic feature (e.g., declining prosody for yes/no questions versus open questions). Since there is no automatic tool to detect questions, researchers need to rely on manual annotation. To this end, short clips that are extracted at random from LFSE recordings could be annotated to find questions and classify them into yes/no versus open before analyzing them for their prosody. But if questions are rare, you would need a very large sample to draw robust conclusions. Thus, it may be more efficient to design a laboratory experiment in which many yes/no questions versus open questions are elicited in a shorter session. In sum, LFSE recordings should only be used when they are the best fit for studying the phenomena of interest and for fulfilling requirements of the researcher's anticipated analyses. Although every project is different, we propose that researchers can determine whether LFSE recordings are right for their research question by stepping through the key decision flowchart we describe in the remainder of the paper (**Figure 1**; see a number of example studies in Appendix H).

2.1. Stepping through the key decision flowchart

Section 3 nudges the reader to consider whether new data is necessary to answer their research question. Collecting new data requires the researcher to obtain (at least) formal ethical approval. If approval is difficult to obtain or if project time is restricted, it may be necessary to make do with existing data. We highlight the advantages of re-using existing data, even when approval is feasible and project time is ample. Re-using data can be equally informative and a great deal more efficient. If collecting new data is necessary, decisions about hardware and sampling made prior to data collection have enormous implications for data processing and analysis later on. The latter part of section 3 therefore highlights relevant features to consider when purchasing recording devices, and discusses the use of supplementary measures that can be combined with LFSE recordings.

Section 4 is targeted at those who have obtained LFSE recordings, either through their own data collection or from an existing corpus. Gathering or otherwise obtaining recordings is the fastest and easiest part of creating a corpus. But annotation is what allows a corpus to become (re-)usable because it converts raw data into analyzable information. In section 4, we give an overview of the kinds of automatic analyses that can be carried out, how well those automated tools are known to perform, and what limitations are already known in terms of participant language and age.

Section 5 provides recommendations regarding manual annotation. As will be explained, even those who hope to use automated tools will still likely need to annotate at least a small portion of their recordings. We discuss the utility of pilot studies and literature reviews to anticipate the required type and amount of manual annotation to achieve the researcher's goals. We then give

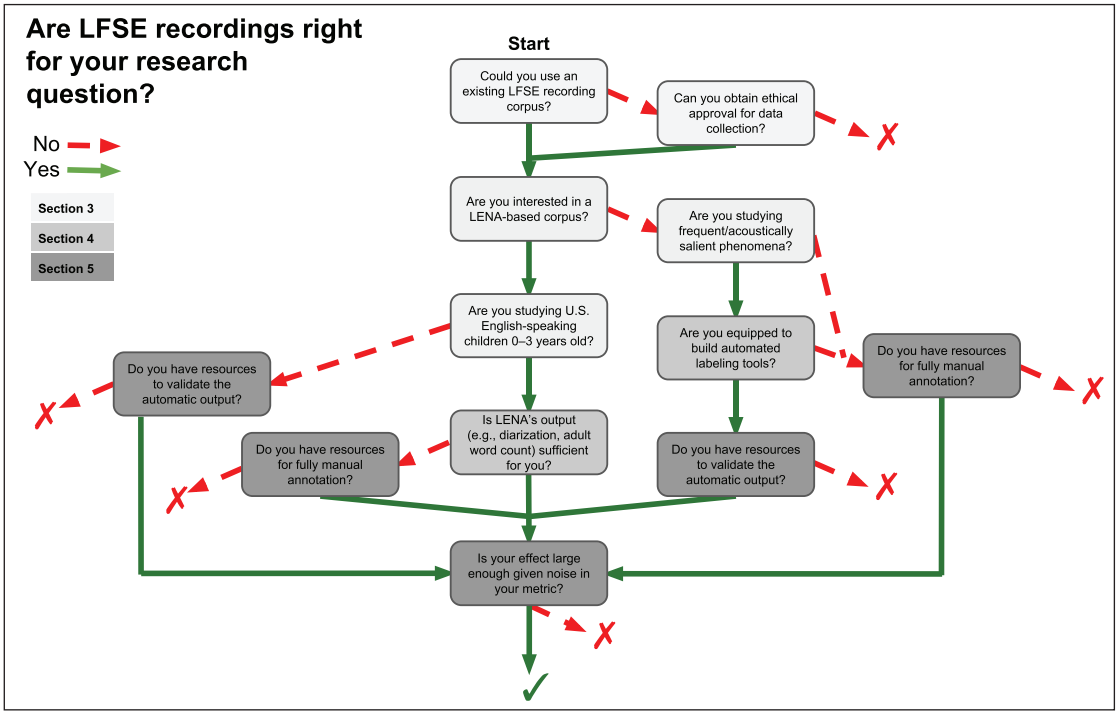

Figure 1: Flowchart of key decisions for those considering a study with LFSE recordings. Those researchers whose path ends with an "X" should instead consider non-LFSE approaches.

basic guidelines for sampling recordings and provide an overview of options available for annotation. We end this section with a brief note regarding reliability and power.

3. Acquiring data

Those interested in using LFSE recordings to answer their research questions should start by first looking into whether existing databases might already satisfy their needs. Working with existing data saves researchers significant time and money, often increases the quality and comparability of shared resources, and poses minimal increased risk to participants, compared to collecting new data.

3.1. Accessing existing data

The advantages of re-using existing data are numerous. First, one avoids the investment of time required to learn about (and put into practice) the laws and ethical procedures for collecting LFSE recordings. Second, by re-using an existing dataset, one can often build on top of previous annotations. Third, investing annotation time in a recording that has already been annotated for information different from the topic under study (e.g., adding gesture annotations to transcriptions of verbal behavior) opens the door to additional analyses that combine both information sources. Fourth, large datasets are typically rife with small mistakes (e.g., typos, unmarked sections where the target participant took the device off). These small issues can be found and addressed as a natural part of data re-use. If there is already an existing dataset that has the right participant sample and recording methods for one's research question, re-using data can be the ideal approach. Naturally, there are many cases where re-use is not possible, for example, if the research question involves participants with Alzheimer's and there are no existing datasets on this population.

There are an increasing number of datasets, including ones with LFSE recordings, available for re-use via specialized scientific archives, or collaboration with the individuals who have collected them. Regarding the former option, one place to find ready-to-re-use LFSE recordings is HomeBank (homebank.talkbank.org; VanDam et al., 2016; HomeBank, 2018). HomeBank is part of the TalkBank framework (talkbank.org; MacWhinney, 2007), which hosts an enormous collection of language recordings and annotations (e.g., CHILDES, AphasiaBank, and many more). On HomeBank, users can find not only raw recordings, but also any automatic or manual annotations that exist. For the time being, HomeBank contributors are mostly developmental language researchers. The participants wearing the recorder are therefore usually children. That said, the creators of HomeBank have made it clear that adult LFSE data are also welcome at their repository, and so we expect that other participant groups will also be represented in HomeBank's archives in the future.

HomeBank data are shared at different access levels (see section 3.2.4 for more details). If readers visit the HomeBank webpage (homebank.talkbank.org), they will be immediately able to download and inspect excerpts that have been approved for public distribution over the Internet. At the time of writing, the fully public section of HomeBank contained one 14-hour, fully transcribed LFSE recording, 159 five-minute fully transcribed segments extracted from recordings from 53 families (3 segments per family), and four daylong recordings (audio only; no annotations), all from American English children (VanDam, 2018a; 2018b; Fausey & Mendoza, 2018). At the same time, the members-only section contained around 5,000 hours of audio gathered from approximately 300 children (primarily acquiring North American English).

Some existing LFSE recording datasets are not included on HomeBank because the researchers who collected the recordings did not ask for the ethical approval to share their recordings at the start of their project. These researchers cannot therefore deposit their recordings on a shared, secure archive. Nonetheless, another mechanism for data re-use is to directly contact these data holders as potential collaborators on specific projects. For example, readers interested in child language and/or social development can reach out to members of the DARCLE (DAylong Recordings of Children's Language Environments, darcle.org; DARCLE, 2018) network, many of whom are data holders open to collaboration on new projects. In general, professional organizations that are focused on naturalistic language recordings (such as DARCLE) may be a good place to find non-shared existing LFSE datasets.

In the long run, however, the goal for language scientists collecting LFSE recordings should be to maximize future data re-use. It is therefore critical for researchers planning to use LFSE recordings to get ethical approval for (eventual) sharing, a topic covered in the next section. Readers leaning toward re-using existing data can now, if desired, skip ahead to section 4 to learn about next steps for annotation and analysis. The remainder of Section 3 addresses the decisions involved for new data collection.

3.2. Preparing to collect new data

Before collecting new data, researchers should carefully consider how to get and implement ethical approval for their planned research and how and where to legally store and back up the data.

3.2.1. Ethical approval and research permissions for data collection

Ethical considerations can be complex when it comes to recording natural language environments. As with other instances of human participants data collection (e.g., behavioral experiments), privacy and informed consent are of the highest priority: Participants must understand their rights and have the ability to withdraw from participation at any time, even just momentarily (e.g., during a private conversation) or retroactively (e.g., request deletion of a clip). However, in the case of most long-format naturalistic recordings, it is impossible to get informed consent from every additional person whose voice is captured on the recordings. The participants are likely to interact with others in both public and private settings during their recording period. How can we cope with the lack of informed consent in these cases? One solution may be to purposefully use microphones that are not very sensitive or are strategically placed so as to primarily pick up the

target participant's voice (e.g., Mehl et al., 2001), though this method does not work for researchers interested in data from the target participant's interlocutors.

Even after having obtained consent from all of those whose voice is recorded, there are a few further issues to reflect on. Because of the long-format nature of the recording, the participant and others around them can forget or simply not realize that events like bathroom visits, spousal arguments, and illegal activities are being recorded. LFSE recordings cannot typically be fully anonymized given the recorded voices and content of talk, so how can we ensure the privacy of those who appear in the data? These are complex ethical questions that do not have any clear answers that can be applied to every data set. Here, we merely note a couple of common solutions.

One solution to maintain participant privacy is to limit the depth of the measurable linguistic information. For example, researchers can decide to only store and allow re-use of derived data (e.g., only speech onset and offset times from different speaker types). Another solution for those working with the audio directly is to use clips so short that the chance of hearing sensitive information is smaller. For instance, Mehl (2017) reports that a 30-second clip with speech typically only contains one or two complete utterances, with fragments of the preceding and following utterances. Therefore, even for the target participant being recorded, anonymity and privacy are maximized by a lack of conversational context in these short clips. That said, automated tools for recognizing voices and words are becoming more sophisticated, so researchers should, generally speaking, still remove identifying information from audio (and transcription) data before considering sharing (Mehl & Pennebaker, 2003; VanDam, Warlaumont, MacWhinney, Soderstrom, & Bergelson, 2018).

When LSFE recordings are made abroad, for example, in traditional/indigenous, non-industrial, developing, and/or low-literacy communities, the ethics become more complicated still. The process of getting permission for participation and later data uses, ensuring participant comfort, compensation for participation, and disseminating information about the findings have to be adapted to each community visited. Researchers must be sensitive to the ways in which participant risk needs to be evaluated in the communities where they want to work. For example, researchers who are working in small communities must design their participant compensations so that they do not destabilize local economic or social relations. Decisions of this type require familiarity with the community before making a plan for how to do the research (see more in Appendix A). With respect to privacy, participant expectations about who can access the data might vary. In some cases, participants might be proud of their (or their child's) participation and request that the data be shared and labeled with their name. If legal guardians are making this decision for a minor, the researcher must consider what happens to the recordings when the child becomes an adult. In other cases, participants may want to limit the kinds of people who have access to their data. For example, in one case we know of, community members were enthusiastic about the wider world seeing their data so long as the neighboring communities did not also get access to it—to best fulfill this community's wishes, the researcher ultimately had to keep the data private.

We cannot assume that participants' conceptions about privacy and data-sharing map onto our own, even if we can achieve perfect translations to the language used by the participants. Cultural sensitivity and direct communication during the process of data collection, annotation, analysis, and dissemination are fundamental for extending the spirit of the ethical guidelines we as researchers design for our own communities. It is therefore crucial that researchers working with populations other than their own are knowledgeable about what is acceptable in those communities and can work in tandem with their local institutional ethics committees to develop a research plan that is designed to effectively minimize risk to participants (see Appendix A for further information).

3.2.2. Ethical and legal issues surrounding personal data

Even if one is not considering sharing recordings with others, it takes a long time to add annotation (particularly for LFSE recordings). It is therefore crucial to think about long-term storage, even before collecting the data. This plan must already be laid out and assessed prior to receiving official ethical approval, and then also explained when gathering informed consent from participants. Data archival is also typically time-intensive, so researchers should consider early on how they can pre-allocate time and monetary resources to make sure it can be done properly (see Meyer (2018) for an excellent review).

Researchers must also consider the laws of the nation, state/province, and city where the data are collected and stored, as well as those of the participants and people in their environment, regardless of where the researcher is residing or institutionally based. Ethical review boards may not be up-to-date on the legal restrictions governing the storage and sharing of digital audio recordings and derived data—an area of governance that is quickly evolving. It is up to the researcher (and their institution) to become familiar with the relevant laws. For instance, Europe has recently put into effect the General Data Protection Regulation (GDPR 2016/679; EU GDPR Information Portal, 2018), providing European citizens with rights over their personal data. This applies when a European citizen's data is being collected or used in any country, even outside of Europe. Sometimes, national regulations regarding the collection and storage of personal data can conflict; for example, a local regulation may ban recordings in public places (such as a supermarket) altogether, whereas a state or provincial regulation may take the stance that behavior (including vocal behavior and thus audio recordings) in such sites is considered public and thus no specific recording permission needs to be obtained. Regulations can sometimes also be at odds with current practice and/or completely impractical, for instance by requiring that anyone who is recorded have the right to request that the recording be destroyed (which requires informing every passing shopper whose voice is recorded about how to contact the researchers). In short, researchers are

responsible for knowing what their obligations are and making a comprehensive plan for how to deal with them.

The process of gaining sufficient familiarity with legal constraints so that one can operate within them can be lengthy. A cost-effective solution can be to collaborate with someone who is well informed about the relevant privacy regulations or who has navigated these intricacies in the past. In all cases, it is a good idea to bear in mind the precise research goal and the minimum risk necessary to address it. For instance, if the research goal only involves global aspects of language use at work, the researcher can altogether discard from consideration any section of the recording that takes place outside of the workplace (e.g., while the participant is in transit or at home).

3.2.3. Data storage and backups

Once researchers are familiarized with the restrictions on data collection and storage, they must decide how to actually implement their storage systems. The simplest option for short-term data storage and backup is to keep recordings and annotations on external hard drives. We recommend that researchers keep multiple, version-controlled copies of their data (i.e., keep back-ups that are tracked for the date and changes to file content). Some modern back-up systems feature "parallel hard drives" (see, e.g., the Lacie 2Big, Seagate Plus Hub, or similar) which allow users to perform backups onto two mirrored volumes simultaneously. Many of these devices also automatically check the stability of the hard drives and alert the user if a drive needs replacing, which can be done easily and inexpensively. If there is a fire or flood, however, the data will be lost if the two hard drives are in the same location. Therefore, extra copies stored in a separate, distant location (e.g., with a dedicated server) are valuable additions to this basic scheme.

The best option by far for remote backup is to use one of the many scientific repositories currently available (see section 3.2.3.1 below). Not only do they allow the data collector to securely and easily store and access their data, but also they facilitate sharing the data with collaborators and, potentially, the wider community. However, as discussed above, special laws may apply regarding where data can be stored and how it can be shared. Many of the scientific repositories we discuss below are located in the United States, where the government is legally able to inspect private servers. Storage of participant data on these United States servers may therefore require explicit permission from participants, special registration with national oversight committees (e.g., the National Committee for Informatics and Freedom in France), or may simply be forbidden. The issues that arise around data storage are complex, and it is our experience that, when possible, consultation with legal experts, institutional staff dedicated to legal compliance, and current LFSE recordings users (via DARCLE) are invaluable during project planning.

3.2.3.1. Scientific data archiving options

What scientific repositories are available for LFSE recordings, and accompanying annotations? Three that are often used are listed in **Table 1**, in order of least to most specific. All three repositories are free and managed *by* researchers, *for* researchers.

The **Open Science Framework (OSF)** aims to provide a home for any and all research output, at all levels of development, from inception to completion. As a result, it has a wide range of capabilities that are eminently useful when sharing data and annotations with others. Since no structure is imposed on OSF projects, it is extremely easy to store data and annotations for the data producers. But, unless the data producers put in a lot of work to clearly organize, systematically tag, and document their data, this freedom of file structuring makes OSF relatively difficult for data re-users. For example, there is no search function allowing one to find datasets bearing on participants of a certain linguistic group or age, or annotations relevant to, e.g., talk addressed to familiar interlocutors vs. strangers.

On OSF, one or more people can be listed as contributors to a project; contributors are administrators who can give other visitors to OSF the ability to view existing files, add new files, or both. Sometimes contributors are not authors (e.g., a lab assistant who is using the platform

Table 1: Three options for scientific repositories.

Repository name	Project file Formatting	How to update data	Data access by non-curators	Other features
Open Science Framework (osf.io)	No specific requirements for files or project structure	Via browser	A choice of: None (completely private), Invited people can read (and write), Anyone can read	Plugins for software such as Google Drive, GitHub (GitHub, 2018), and others; Storage in USA or Europe
Databrary (databrary.org)	Some aspects of project structure specified	Via browser	A choice of: None (completely private), Invited people can read (and write), Anyone can read	Data annotation with Datavyu software (Datavyu Team, 2014), some APIs
HomeBank (homebank.talkbank.org)	Project structure and file structure must follow one specific format	Through personal contact with HomeBank personnel	A choice of: None (completely private); Any HomeBank member can read; Selected HomeBank members can read; Anyone can read	Data annotation and analyses with CLAN software

to manage other data annotators). In these cases, non-author contributors can be tagged such that, when a DOI-based citation of the project is generated, their names will not be acknowledged. Anyone with a web browser can interact with the OSF framework, which is very light and, in our experience, works well even in places with low-bandwidth Internet connections. Newly created projects are automatically set to be private (i.e., are not searchable and not accessible by users other than the collaborators listed). Contributors with administrator rights can generate view-only, anonymized links of parts or all of the project, which is useful when sharing stimuli, data, analyses, and supplementary materials during peer review. Researchers can also set their projects to be, partly or wholly, publicly accessible, at which point anyone can visit it on the web. The fact that some parts can be public and some parts private is useful because it allows data holders to open some parts of their dataset to the public (e.g., anonymized or secondary annotations), but still limit access to other sections.

OSF also offers a number of Application Programming Interfaces (APIs) and plugins (i.e., bridges to and from other systems), which has contributed to its increasing popularity among researchers. For instance, data holders can link their OSF project to an already-existing folder on other cloud-based data storage systems (e.g., Google Drive, Dropbox, or OwnCloud). A direct link between storage systems makes data transfer much easier. Finally, OSF is the only repository among the three we discuss that allows users to choose where their data are stored (USA, Canada, Germany, or Australia).

Databrary presents itself as a repository for primary and secondary data, particularly on child development, and with an emphasis on multimedia files. As a consequence of their focus on child development, data holders are required to define certain properties of their data that are thought to be crucial, such as child age, country, and ethnic background. This results in a distinct advantage for data re-users interested in development: one can perform searches for these features. However, the clear disadvantage is that the repository is less likely to be relevant to those who are not interested in child development.

The creation of projects and data upload/download in Databrary are about as easy as they are in OSF, and options for sharing are equally rich and diverse, including the option of releasing some files publicly but not others. Two unique features of Databrary make it particularly appealing for sensitive data. The first is that in order to share one's project with someone, that person must be authorized through Databrary's process (or under the supervision of someone who has access rights). The process of gaining direct authorization involves Databrary communicating with an official in the individual's organization, who is made aware of the importance of appropriate legal and ethical handling of these data, and the fact that the institution is legally responsible if that individual (or anyone under their supervision) commits any infractions (Databrary, 2018). Although it is easiest for Databrary to interact with American organizations, most of which have a local Institutional Review Board (IRB) which will be able to handle this request, they have historically made the extra effort to adapt their process for foreign organizations. We have not had the opportunity to assess what would happen for an independent researcher (i.e., one not affiliated with an institution). The second feature that makes Databrary a promising archive is that access to a repository is only granted when that person specifies an end-date for access. This apparently minor detail makes it easier for data holders to keep track of who has access to their data at all times, and serves as a reminder to delete or dispose of copies of the data when the collaboration comes to an end. As a final note, the system of APIs and plugins is not as developed in Databrary as it is in OSF.

While OSF and Databrary include a wide variety of types of data, **HomeBank** is part of the TalkBank framework and is thus centered on language specifically. Researchers using language data benefit when using HomeBank because its data contributors are encouraged to use a single, widely shared coding scheme for making linguistic annotations (CHAT, MacWhinney, 2000). The CHAT coding scheme specifies formatting requirements for making transcriptions and annotations of sounds, words, sentence structures, and more. While this process creates more work up-front in creating annotations, it greatly facilitates analysis. Contributed data, even across corpora, are analyzable with a single script (e.g., in the programming language of the analyst's choice) or with the free software provided by TalkBank to help researchers to perform many useful frequency and co-occurrence analyses when contributed data contain transcriptions (CLAN, MacWhinney, 2000). It is also possible to contribute video or audio data with no annotations. HomeBank does not have a user-based system like that of OSF and Databrary. Therefore, unlike OSF and Databrary, data holders cannot directly interact with their data on the site: the HomeBank team handles all contributions and annotation editing (https://homebank.talkbank.org/contributing.html).

The absence of a user system also changes the options for sharing because it is not possible to share files directly with individuals. Instead, HomeBank proposes four basic sharing levels. The most restrictive is Embargoed, in which case the data are available only to the HomeBank team and the data contributors. Contributors may opt for this access level if they are only looking for storage and safekeeping, or if they are still preparing their data but would need to first check that they comply with CHAT formatting. In the latter case, data holders may ask that the embargo is lifted at a certain date, or when an event occurs (e.g., a paper describing the dataset is published, or when the data contributor passes away). The next level of sharing is Members, meaning that recordings are available only to the HomeBank members; "members" are researchers who have been vetted in a one-on-one interview with a representative from the HomeBank team. During the interview, conditions for data re-use and ethics are discussed, and applicants submit a signed data re-use agreement acknowledging their understanding of the rules. There is also a special section of HomeBank called "Sensitive", intended for sensitive data (e.g., involving vulnerable participant populations). Data within this section are only available through a special password that members need to apply for separately by describing their

need for access and gaining approval by the data holder and the HomeBank administration team. The final and most inclusive level for sharing is Public, i.e., accessible by anyone on the Internet. Notably, the basic metadata for all the collections, including the Public, Members-only, and Sensitive databases, is visible on the repository overview (https://homebank.talkbank.org/access/) for those considering data re-use.

In our view, one limitation of all three repositories pertains to version control for annotations and the structure of data curation. All three repositories assume that someone will be in charge of the data and will update it as, for instance, more annotations become available. But it is not clear how these different versions of the annotations should be stored and made available. We explain this issue in more detail and provide some solutions in Appendix B.

In sum, researchers have a great deal of options to consider in finding the optimal way to legally, ethically, and practically store and share their LFSE recording data. Once this is settled, a next consideration is what hardware and software they should use to carry out the project.

3.3. Collecting new data

Decisions regarding what recording device to use (i.e., the hardware) should be related to decisions about the kinds of analyses that will be performed which, in turn, relate to the population being studied. In simple terms, readers working on American English children aged 0–3 years of age should strongly consider using the LENA device paired with the proprietary LENA software because that set of tools has been developed targeting their specific population. We will introduce LENA's hardware and, in broad strokes, the measurements its software provides in section 3.3.1; a fuller evaluation of the LENA software performance in American English learners and other populations is provided in section 4.2 and Appendix C. Readers interested in other languages and age groups may also consider the LENA system, but should only do so with the understanding that the measurements it provides may be inappropriate or inaccurate for their use case. We therefore also review alternatives for the hardware in section 3.3.2, and for the software in section 4.3 below.

3.3.1. LENA

Only one piece of equipment comes with pre-processing software that works out of the box: the LENA. We will provide here a very short overview, since there is a host of information on the possibilities and limitations of the LENA products elsewhere (see, e.g., Gilkerson & Richards, 2008, for an introduction and Ganek & Eriks-Brophy, 2018a, for a review of published research using LENA). In a nutshell, LENA's hardware and software are durable, stable, and simple to use by all participants involved in developmental language studies (researcher, practitioner, parent, and child). We very briefly describe what the LENA system is like in the following section, but we recommend that interested readers check out Ganek and Eriks-Brophy (2018a)'s excellent review of research that has been carried out with LENA. It provides an overview of the variety of ways in which researchers have harnessed the strengths of the LENA hardware and software.

The LENA system's automated estimates are most reliable if the wearer is recorded for 12 or more consecutive hours within a single day, though the device can accommodate recordings being split over several days (Xu, Yapanel, & Gray, 2009). Once the recording is done, the researcher/practitioner/other user connects the recording device to a Windows computer with access to LENA-licensed software (either locally or via a web browser) that automatically extracts and analyzes the recording. At the time of writing, there are two licenses ("Pro" and "SP") allowing users to keep a copy of the recordings. The Pro and SP licenses differ crucially in that the former does all speech processing in the local machine, whereas the latter uploads recordings to a remote ("cloud") storage location where they are processed, a set of statistics is returned, after which the original files are deleted. Although there are no plans to discontinue the Pro license, there is also no budget for updating it as Windows systems change, meaning that it will become more difficult to maintain the necessary hardware and software needed to use the Pro system with time. If improvements are made to LENA algorithms in the future, Pro users may also not have access to them. In contrast, the SP license accesses LENA's software via an Internet connection—either from a computer on which LENA Hub software is installed or through a web browser such as Safari or Chrome. The ability to securely upload, process, and inspect results from recordings from any Internet-connected computer may be an important advantage for studies spread over multiple locations.

The LENA system provides several types of automated annotations. For both Pro and SP licenses, users are given a file in which the audio signal has been classified into the following classes: key child (wearing the recorder), other children, female adults, male adults, TV and electronic sounds, overlap (between any of those categories), and silence. For each segment tagged as belonging to the key child, there will be an estimate of what portions are vegetative (e.g., burps) or crying, as opposed to speech. For each stretch of speech tagged as being produced by an adult, there will be an estimate of the number of words spoken in that stretch. At a second level, the LENA software also derives a few descriptive statistics that are averaged over five minutes, one hour, and the whole day: Child Vocalization Counts (CVC), Adult Word Counts (AWC), and the total number of turns involving the key child and an adult (Conversational Turn Counts or CTC). In addition, thanks to the large-scale norming study carried out by the LENA Foundation (Gilkerson & Richards, 2008) the system also provides an estimate of child vocal maturity compared to other children of the same age and sex using a standardized score called Automatic Vocalization Assessment (AVA). All of the above will be most accurate for children learning American English aged between 2 months and 3 years, which is the population used to train and test the analysis software (Gilkerson & Richards, 2008). As with any tool, when one deviates from the population on which the tool was developed

and validated, the researcher should take care to test the extent to which the measurements are still reliable (see section 4.2 for more details on validation).

Researchers may feel hesitant to use LENA's recorder and software system because of its relatively high price (12–15k US$ at the time of writing). Consider, however, that 15k US$ buys approximately 1500 assistant work hours (assuming a minimal $10/hr). Given that it takes, in our experience, at least a 1:7 ratio of audio minutes to annotation minutes for preliminary transcription with coarse utterance-boundary placement (often longer; see Section 5 for manual annotation), the researcher would get ~215 hours of annotated audio for the same price, not including assistant training time. Projects with more than 200 hours of planned audio analysis may therefore be cheaper with LENA, so long as the LENA output is suitable for the research question without further annotation. Another consideration is that investing in alternative processing pipelines would likely be very expensive. For example, the LENA Foundation spent four years and millions of dollars researching and developing their system (Gilkerson & Richards, 2009: 6). The researcher or practitioner who wants to "strike out on their own" might have to re-do all of this if starting from scratch, likely ending up with results that are less reliable than LENA's.

A final point is worth repeating: researchers and practitioners across the language sciences are likely to be interested in language properties that are different from those provided automatically by the LENA system, including, for example, measures of lexical diversity, syntactic complexity, and who is being talked to. We do not know of any automatized system currently being developed that includes these measures, and doubt that they will be available in the near future. After years of collaborating on the development of automatized tools, we can confidently say that automatically transcribed speech from LFSE recordings will not be available for at least the next five years.

3.3.2. Others

LENA products are less appealing when the derived measures are not relevant to the research question (e.g., the study is on music exposure or adult participants). Some potential users may also decide that the performance is unlikely to be good enough for their research purposes, for instance, when a previous validation study on one's language or population of interest revealed levels of performance below what is necessary to estimate the effect of interest. Or, for example, manual inspection of a recording may reveal that the audio quality of the recordings is not high enough for one's needs (e.g., for a researcher interested in the detailed acoustic properties of sounds like 's' and 'sh' in English).

There are many alternative recording devices for LFSE recordings—any wearable, lightweight recording device is a candidate. The best choice for a project, as always, will depend on the research question and population. We provide here a few examples of studies not using the LENA system (device + software). Some researchers have opted for the LENA hardware even if studying adults (e.g., the "Prof-life-log" corpus; Ziaei et al., 2015) because it has FDA approval and is both sturdy and easy to use. Other researchers, notably the proponents of EAR (Mehl, 2017), have opted for iPods, which are just as lightweight (though not as sturdy) as the LENA, and use a platform that can be programmed. Others have come up with solutions that work well only in the home, such as recording systems coupled with wireless microphones (Wells, 1979). One can also simply use lightweight recording devices that are already favorites among field linguists (e.g., Olympus handheld recorders; Casillas et al., under review) or even "spy technology" like USB audio recorders that have a battery life of 15 hours (Scaff et al., in preparation). This last option is cheap and extremely lightweight, though it is more prone to equipment failure than the other options, in our experience.

When considering which hardware to buy, we recommend that users pay attention to aspects such as sampling frequency (higher is better), bit rate (higher is better), compression (lower is better); as well as other features that are harder to find in technical specifications, such as the faithfulness of the frequency response (flatter is better), and, for fieldworkers, how stable its performance is given variation in temperature and humidity, and how easily one can recharge the recorder. One should also take into account hardware limitations such how many hours of recording can be captured at the desired recording settings (e.g., high bit rate, low compression) on a single battery charge and with the device's memory limitations. The simplicity of user interface is also an important consideration for those allowing their participants to control when the recorder is on and off. For example, those studying aging individuals should consider how easy it is to remove the recorder, stop it, restart it, and put it back into place given the variety of mobility restrictions faced by participants who are expected to control their own recording times.

Whatever the hardware, one general recommendation is to try to use clothing where the device is tight-fitting, so as to avoid noise from the device dangling or scratching against the wearer's clothing. Another general piece of advice is to place the recorder in the outermost layer of clothing possible. Thus, under cold-weather conditions, recording set-ups that can be easily clipped onto jackets, sweaters, and t-shirts as needed may be the best choice. However, adjustable setups of this sort depend on users remembering to move the device from one position to another, which may tamper with ecological validity.

3.4. Combining LFSE recordings with other measures

Our discussion has so far only focused on audio recordings, but many researchers using LFSE recordings are interested in non-verbal communication. For these readers, we strongly recommend surveying recent developments in the field of "life-logging" or "quantized self" wearable devices. For instance, Casillas, Brown, and Levinson (2017, under review) paired audio recordings with a small camera that can take photos at a fixed rate (e.g., every 10 seconds; the Narrative Clip 1), and fitted

it with a fisheye lens for a 180° front view from the wearer's perspective. These two data streams (audio and photos), provided by two off-the-shelf devices, are both time-stamped and can be subsequently combined into a video. In their study, the photos helped the researchers not only to assess the variety of environments participants were in, but also to resolve difficult talker identification cases (e.g., when several interlocutors sound alike). Even better is when the data streams can be integrated within a single device, as achieved by Abels (e.g., Abels & Vogt, 2018), who combined audio recordings with heart-rate measurements to have a psychophysical correlate in the wearer, as well as radio-frequency emitters/receptors worn both by the target participant and others in the community. The radio system allowed her to estimate the distance between the wearer and potential interlocutors. These technological developments are exciting and we welcome readers who have explored other add-ons to discuss their work in professional circles around LFSE recordings, such as DARCLE. Importantly, we should warn readers that hardware development and adaptation is costly, so it may not be ideal for practitioners or scientists who are hoping to gain quick insight into natural language use. Adding more types of data (e.g., audio only vs. audio and images) may also require researchers to make different decisions regarding participant consent and privacy, secure storage, and data sharing.

4. (Semi-)automatic analyses

In this section we will give a basic overview of automated analyses for LFSE recordings, including descriptions of the various tools used by the LENA system, a summary of their accuracy scores, information about alternatives to LENA, and a description of some limitations to alternative (non-LENA) tool development.

4.1. Basics of automatic audio analysis

Before jumping into our discussion of automated analyses, two caveats are in order. First, in what follows we will explain the processes involved in automatic analyses as if they were a strict sequence of stages. This is, in fact, inaccurate for some tools (e.g., Wang, Neves, & Metze, 2016 do voice activity detection and noise classification in a single step, not as two sequential steps). However, we limit this guide to describing the overall performance of these systems, and not the details of their implementations. The second caveat is that the only fully established, fully automatic analysis software available to researchers at the time of writing is LENA's. Therefore, in this subsection all of our examples are from LENA. We introduce current alternatives to LENA and other software under development later in this section.

Any pipeline of automatic analyses applied to LFSE recordings will likely have some or all of the following stages, applied over brief (e.g., 0.1 second) stretches of the audio signal:

1. Is this stretch a vocalization or not? In the speech technology literature, this is called *voice activity detection* (or *speech activity detection*; the difference between the two being that the former, but not the latter, includes non-speech vocalizations such as yawning and crying).
2. If it is a vocalization, who produced it? This task is called *talker diarization*. In many systems trained to automatically process LFSE recordings, this question is more likely to be resolved at the role or category level (i.e., "which type of person?") and not at the individual level (i.e., "which person?"). That is, the systems we are familiar with will return "participant" and "mother/doctor" (or sometime even just "child," "female adult," and "male adult") and not "Robert" and "Anne".
3. If it is not a vocalization, what is it? The LENA system returns the following alternative labels: silence, electronic noise/TV/radio, overlap (between any two sound categories, for example target child and adult speech, target child and radio, etc.). Other, non-LENA, noise classifiers can return more categories (e.g., Wang et al. 2016 classification includes singing, engine sounds, and nature noises, among others).
4. If it is a vocalization, what are its characteristics? Systems diverge the most on this point. For instance, the LENA system estimates whether a stretch is speech or non-speech for vocalizations from the *target child*, but estimates the number of words spoken for vocalizations from a *female adult* or *male adult*.

Often, users do not want to analyze each individual stretch of audio signal, but would rather extract global statistics (e.g., average minutes of speech per hour) from the full recording, sometimes even at the participant level from several recordings. This type of analysis is typically done by aggregating over local classifications from individual stretches of audio. For instance, the estimate of adult word counts in LENA is, quite simply, the sum of all the local estimates (e.g., words in individual stretches of vocalization) for that day. It should be borne in mind, however, that summing over local classifications may under- or over-estimate global quantities. For example, an automated analysis that is based only on near and clear speech will not count all speech events (it will miss all the far away or whispered speech), and thus lead to under-estimations. A system that is trained with a very talkative set of individuals may yield "false alarms" (i.e., report that the wearer is talking, when in fact they are not) when used with a less talkative population, leading to global counts that are too high.

A final important point is that LENA's software has been developed by researchers interested in child development, and thus performance might not generalize to populations different from the type of children and recording settings the system was designed for. For instance, one can anticipate a decline in performance when the LENA is used in a daycare, with groups of children older than age three,

or when the recorder is worn by an older adult. That said, researchers have successfully used LENA in these settings; see, for example, Soderstrom and Wittebolle (2013) for a daycare study, Jackson and Callender (2014) for a school study, and Li and colleagues (2014) for a study on aging.

4.2. Accuracy of LENA automatic analyses

Both the LENA Foundation and independent researchers have evaluated the accuracy of LENA estimates at several levels, including talker diarization (determining who speaks when) and the three main global measures provided: adult word counts, child vocalization counts, and conversational turn counts, each of which is summarized below. In addition, some researchers have built pipelines that derive additional metrics from LENA's automatic annotations (see, for example, the HomeBankCode repository, HomeBankCode, 2018, on GitHub, GitHub, 2018). We will not review the accuracy of such derived metrics here.

4.2.1. Talker diarization

Most previous work evaluating talker diarization has done so by presenting audio clips to listeners and asking them to classify the clips into the same categories that are provided by the LENA system (e.g., Bulgarelli & Bergelson, under review; VanDam & Silbert, 2016; Xu et al., 2009). This apparently straightforward approach glosses over two important complications. The first occurs when a clip of audio contains speech from two different categories (e.g., target child and female adult). If the human can indicate that both categories are present, then the stretch might be counted "correct" whether the system identifies the clip as belonging to one category or the other, boosting performance artificially. A second, related point is that typically vocalizations also include background noise or short stretches of silence, which means that the process of segmentation was not perfectly precise. Both of these issues are avoided by using a different accuracy metric, which we will introduce in section 4.3. Setting these issues aside, current reports suggest LENA has very good performance for American English learning children (percent of true cases identified, "recall": ~75–85%; percent of the identifications that were accurate, "precision": ~75%) and somewhat lower for children learning other languages (with variable results; see Appendix C.1). However, many of these evaluations have collapsed judgments of who is speaking across the target child (wearing the recording device) and other children (i.e., any other child). This approach only makes sense when the target child is the only child present but would not be advisable for settings with multiple children present. Finally, we know of no reports on LENA talker diarization accuracy when the hardware is worn by adults.

4.2.2. Adult Word Counts, Child Vocalization Counts, and Conversational Turn Counts

As mentioned above, the LENA system estimates the total number of adult words heard by the child over the recorded day (AWC), the number of target child vocalizations produced over the recorded day (CVC) and the number of conversational turns (i.e., exchanges of talk between two or more speakers; CTC) between the target child and an adult. The current practice for evaluating these measures is to extract clips (typically five minutes in length) from the recording and play each vocalization identified by LENA for a human annotator. The annotator then adds a manual "gold standard" answer against which the automated counts can be compared. For example, for AWC, one can transcribe the content of each utterance (manually or by clearly repeating the speech contents into a speech recognizer), count the words they contain, and then compare the estimated word count against the human word count. Most often, researchers check for reliability at the level of the entire clip, reporting the degree of similarity (e.g., with a correlation test or an estimation for average error rates). Further details are provided in Appendix C.2 but, in a nutshell, performance for AWC and CVC are good (correlations above .6) both when LENA is used on American English and other languages. One exception is conversational turn counts, for which the average correlation over three reports is lower than .3. Again, to our knowledge, no studies have yet validated these measures when the hardware is worn by adults.

4.2.3. Other measurements provided by the LENA software

There is less information on the accuracy of other estimates, such as the classification of child segments as being speech or non-speech. For these, the LENA Technical Report (Xu et al., 2009) provides accuracy estimates of 75%, and 84% respectively (see Gilkerson et al., 2015, for a report on Mandarin-learning infants, Elo, 2016 on Finnish infants).

4.3. Available alternative systems to LENA

In this section we will review currently available alternatives to the LENA system for generating automated annotations. We will focus on two recommendations for processing full-length LFSE recordings: KALDI and DiViMe. Both require some basic use of the Unix command line (see, e.g., online courses such as The Unix Shell). The most flexible and powerful platform is KALDI (Povey et al., 2011), which in the last few years has become a central system for sharing speech technology code. For readers with no prior informatics and speech technology expertise, KALDI is intimidating. Users must be able to understand and respond to error messages while installing KALDI's numerous other required software packages. Moreover, most KALDI tutorials aim to enable the user to build speech recognition systems (i.e., transcription of the audio), and thus users must invest time to cull just the relevant steps for their own work. That said, there are several excellent tutorials available should researchers want to take on this task (e.g., KALDI for dummies, 2018).

Although the there is no real open-source, language-general, and population-general version of LENA, there is a relatively easy-to-use, open alternative being developed: DiViMe (Le Franc et al., 2018; ACLEW/DiViMe, 2018).

As of July 2018, DiViMe contains two alternative voice activity detection routines, a talker diarization system and an evaluation package. Although DiViMe can only be used on the command line, its creators are motivated to make it easy to use by anyone with basic technical knowledge. For instance, all tools are launched with simple commands such as "TOOLNAME data/", with the tool then being applied to all wav files inside the "data" folder. Further information on DiViMe is provided in Appendix D.

4.4. Population-specific training for automated analyses

The tools described above will only go part of the way toward the realization of adaptable automated analysis of LFSE recordings. What we are missing are tools that can be trained on or adapted to the population each researcher is interested in. Indeed, one useful feature of the LENA output is that the target child is distinguished from adults (and, to an extent, from fellow children). To accomplish this, the LENA Foundation created a very large training set (on win which the target child had been manually tagged in recordings of young children growing up in a socioeconomically representative sample of the American population. This training set allowed the LENA engineers to create and test models of the kinds of vocalizations children are likely to produce at ages 2–36 months, thereby allowing their users to have "target child" speech tagged on the basis of both surface features (e.g., "sounds like a child and speaks close to the microphone") and age-related features. As this description hopefully makes obvious, it is unlikely that any individual researcher has the time and resources to train another general purpose, multi-age system to this level of accuracy.

Readers may wonder whether it is truly necessary to try and attain adaptable automated analysis tools for LFSE recordings specifically. Evidence shows that it is: we recently helped organize the "DIHARD Challenge" (Ryant et al., 2018; see also Ryant et al., 2019), a competition in which speech technology teams applied cutting-edge solutions to voice activity detection and talker diarization of conversational speech. We found that all of the systems performed worse on the LFSE recording data than in the other, non-LFSE datasets included in the challenge. The current reality is that speech technology specialists are not targeting recordings of real-life conversations. As a result, they are not developing tools that generalize to the very challenging LFSE recording context. Solving this issue will necessarily require that researchers who collect such datasets share at least some portion of them, so that they can be included in similar challenges to promote further development of tools that can perform well on recordings of this type (see also Schuller et al., 2017; 2019).

5. Manual annotations and subsequent analyses

In this section, we provide advice as to when and how to gather manual annotations. We believe that, for many research questions, it is not feasible to rely exclusively on out-of-the-box automatic measures because most of the relevant speech technology tools are still in their infancy. The possible exception to this warning are studies that are designed to be based on American English learners aged 0–3 years with the LENA system's well-documented outcome measures (e.g., adult word counts). For every other case, including when LENA is used with non-validated populations, some amount of manual annotation is unavoidable. However, this should not deter readers: If one has a clear idea of what the goal is, it may be possible to devise a form of annotation that serves one's purposes and is feasible, despite the large scale that LFSE recording datasets tend to have.

The first consideration is, for one's research question, does the whole LFSE recording matter, or would a smaller sample do? If the whole recording matters, could automatic measurements be estimated and then validated by annotating a smaller sample? And, in both cases, could an automatic method be used to extract the recording clips needed for annotation? Section 5.2 provides guidelines to help find the best sampling method given those questions. During this process it's also worth considering what workflow could be used to manage annotation (e.g., first identify segments to annotate with a tool, then randomly allocate segments for annotation to coders A, B, and C within each LFSE recording, etc.) This topic is covered in more detail in section 5.1. Though we introduce sampling plans and data management here as two separate tasks, we hope that, eventually, there will be freely available data management software that allows clip sampling and distribution to be managed jointly.

In what follows, we start by providing general guidelines on how to manage media and annotation data. This is a problem faced by any researcher with a dataset, but we highlight tips from our experiences with LFSE recordings specifically. We then go over data sampling techniques and annotation software options useful for creating manual annotations in such large-scale recordings.

5.1. Database management

Most people will not annotate full LFSE recordings, but focus annotation instead on short clips extracted from them. That means that, for each full recording, one also typically has a set of smaller sound files with time-aligned annotations. For certain tasks, researchers may want to extract even smaller clips for different levels of analysis, for example, individual vocalizations from multi-minute clips. These desired features lead to a *nested* representation of the data file: vocalizations are found within annotated clips, which themselves are only part of a very long (and only partially annotated recording). This situation can quickly become much more complicated. Imagine that you have done a first pass through the recording to segment out periods with speech (e.g., using LENA's software or DiViMe), and have produced one clip, from which you extract "vocalizations" for trained annotators to transcribe. For some vocalizations, the transcriber might like to adjust the onset and offset boundaries of the annotated vocalization, or break the vocalization up into two parts, or perhaps even assign the vocalization parts to different talkers. Imagine, too, how you could compare different automated diarization outputs to each other and to manual diarization edits made by a human annotator. These actions lead to a second set of annotations that are *not* nested within the first.

It is for this reason that we recommend readers to consider using annotation software that can organize annotation information with the same structure needed by the researcher, and to use that software throughout the whole annotation process. The ideal software, in our view, should handle both (a) the basic general structure of LFSE recording data, i.e., recordings belong to target participants, clips belong to recordings, vocalizations belong to speakers within clips, etc., and (b) its heterogeneous nature, e.g., recordings and clips are audio files, target participants are represented by clusters of metadata, annotations only partially cover the recording duration, and annotations might be stored in different formats or produced by different annotators. To our knowledge, there is only one open source software option that fulfills all these needs and more: the Emu Speech Database Management System (Emu-SMDS; Winkelmann, Harrington, & Jänsch, 2017). Some of the strengths of Emu-SMDS include:

- Database managers define the global architecture, which allows the system to perform checks for well- or ill-formedness of certain annotations;
- It does not assume that the user will perform exhaustive annotation in the file and therefore does not automatically attempt to load the whole sound file onto memory when inspecting a subpart;
- It does not rely on files being available locally, but rather allows users to access them (and annotate them) via a web browser. Thus, one avoids the security risk involved in directly providing copies of sensitive data to non-collaborator participants, e.g., undergraduate or temporary research assistants;
- It does not try to replicate all annotation software features, but is instead interoperable, which allows the user to use a specialized program for specific tasks (such as Praat, Boersma, 2009, for annotations that benefit from a spectrogram);
- It keeps track of changes made to the annotations (using git, Git, 2018);
- It is natively connected with a set of automatic analysis routines which may be useful later on, such as systems that automatically align matched audio and transcription files ("forced alignment"; WebMAUS, Kisler, Reichel, & Schiel, 2017).

One disadvantage of Emu-SMDS is that it requires a significant initial time investment in learning how to use the system, with the biggest hurdle being the need to set up a local server. When one of the authors contacted the current maintainers of Emu-SMDS, they provided tips to help with set-up, and also offered to set up the server for us in their local environment. For readers who intend to request this service from the Emu-SMDS team in order to avoid the hassle of setting up their own server, this is yet another reason to ensure that appropriate data sharing permissions are set up early on in the project (see Section 3.2 for discussion).

Future virtual database managers may also be able to "propose" samples to the human annotator on the basis of the regions or types of recorded data that the system is trying (but failing) to analyze automatically. This type of system, sometimes referred to as "human in the loop" or "interactive machine learning" (e.g., Holzinger, 2016), is not yet available. Of course, the alternative to database management software is to have humans make these decisions. For those of our readers following this more traditional route, we have a few recommendations in Appendix E. However, we strongly discourage readers from this route—investment in a good database management system today will save one much time later on.

5.2. Sampling

As mentioned above, most annotations for LFSE recordings are partial, rather than exhaustive (i.e., all recorded minutes from all participants are annotated). When creating a partial annotation, researchers will need to sample from several higher levels in the design, including participants (out of all recorded participants, which of them will be partially annotated) and LFSE recordings (out of all the recordings associated with a participant, which ones will be partially annotated). If there are participant subgroups (e.g., an intervention versus a control group), even further levels need to be considered. Discussing sampling at these higher levels is beyond the scope of this paper, where we hope to focus on the problems exclusively or primarily raised by LFSE recordings. Therefore, we focus on how to sample in order to create a partial annotation within a single daylong recording.

Many language scientists have opted to extract samples in one of a few specific ways. For instance, in the EAR research, 30-second clips are sampled every 12 minutes (Mehl et al., 2001; 2007; Mehl, 2017; a similar technique is used by Scaff et al., in preparation and Ramírez-Esparza et al., 2014; 2017). We will call this *periodic sampling*, and it is one implementation of *random sampling*. We use the term "random" because there is no particular reason to target these times. Random sampling can also be aperiodic, for example, 5-minute segments from randomly selected, non-overlapping moments in the day (Casillas et al., under review). The advantages of random sampling are that it is easy to implement and that the estimates produced from these samples (e.g., quantity of speech produced by the wearer) are completely unbiased. The main disadvantage, however, is that for certain research goals, random samples may not provide sufficient amounts of relevant data (e.g., in the case of yes/no questions vs. open-questions mentioned previously). Particularly, if the researcher wants to measure specific linguistic properties like specific words used or utterance complexity, many of the random samples may not contain any speech at all and so may prove useless.

A common alternative is *volume sampling*: extracting segments of the recording in which the phenomena of interest are likely to occur. For instance, one can extract 5-minute chunks during which vocalization or word count estimates are highest, in order to focus analyses on the wearer's peak talk for the day. Others favor chunks with high conversational turn counts, so as to target times in which the wearer interacts with others. Demarcating similarly high-speech-volume regions without LENA or DiViMe is, unfortunately, still a manual process for the

time being. It requires the annotator to scan the entire recording for candidate high-volume moments, either by hand or by using a low-level tool like a Praat script that looks for intense acoustic energy on frequencies commonly used in human voicing (e.g., 80–300 Hz). The researcher must then manually select the best segments from that set of candidates. While these approaches for gaining high-volume samples are promising and reasonable from an analytical standpoint, there is no work validating whether findings extracted from such samples match well with more extensive annotations of LFSE recordings or with more targeted and standardized data collection methods (see also Bergelson, Amatuni, Dailey, Koorathota, & Tor, 2019 and Tamis-LeMonda, Kuchirko, Luo, Escobar, & Bornstein, 2017). It is likely that a small and unique five-minute sample from a given day of recording will be more variable across participants than a controlled recording because the at-home samples will vary more in context: while some might happen during mealtimes, others will happen during phone conversations, TV-watching, etc.

With this issue in mind, other researchers have preferred extracting five minute segments during which the ongoing *activity type* is kept constant (e.g., meal times found between 11AM and 2PM in the recordings for all participants, e.g., Mastin, Ellwood-Lowe, Marchman, & Fernald, 2016). Naturally, this is only possible if one already has an idea about which activities are insightful and can also clearly recognize the activities from the audio/video. Processing the data this way is more time-consuming for the user and may also ultimately limit generalization.

Future methodological work should compare these and other sampling techniques against other, more established, methods such as directly eliciting the desired behaviors or using standardized tests. Whereas samples from whole-day recordings are likely to be more variable, we could imagine that they might yield more valid measurements of communicative behaviors than short-scale elicitation tasks or standardized tests because they represent performance in the participant's everyday life. However, if LFSE recordings gave no further benefit over these other measurements, the work put into assessing their validity could help researchers feel certain about the efficacy of the more targeted, controlled recordings in the future.

5.3. Annotation software

If readers are already used to working with a specific annotation software, a first step when thinking about using it with LFSE recordings is to open up a sample file, navigate through the file from start to finish, and try to save some example annotations. If the annotation software performance is choppy, it may make more sense to learn a new system, rather than lose time working with the old one. Appendix F provides an overview of some popular, currently available systems.

Depending on the software chosen, there may already exist a set of training materials for annotations that could benefit the researcher in both the long and short term. By using one of these community-oriented annotation formats researchers can (a) speed up the initial decision-making process of what exactly to annotate and what exact guidelines to use (e.g., how to deal with edge cases when annotating who the speaker is talking to) and (b) ensure that any annotations produced manually are in a format where they can be analyzed and re-used using shared scripts and tools. We highly recommend the DARCLE Annotation Scheme (Casillas, Bergelson et al., 2017), for which at least one community template contains detailed training materials, including a web-based gold-standard test for new coders, in both English and Spanish (the ACLEW Annotation Scheme; Casillas et al., 2018). Readers are also strongly encouraged to read Ganek & Eriks-Brophy (2018a) for an overview of manual annotation systems used in previous LFSE work (including, e.g., the Social Environment Coding of Sound Inventory system adapted by Ramirez-Esparza et al., 2017).

5.4. Final recommendations regarding annotation

One of the most important questions to ask, even before collecting data, is how much annotation is needed to answer the research question. In many cases, there is no easy answer to this question, and whatever answer is given depends partly on the goal of annotation. Will annotation be the primary source of data, or is it done to check the quality of the automatic annotation? For the latter, see Appendix G for an overview of previously used sample sizes. Is annotation carried out to train or re-train an already extant automatic annotation system? This is not a straightforward question to answer either. Power analysis, as used in experimental research, may be a useful approach. They require the researcher to assess how much data will be required by estimating: (a) how large the effect will be, (b) how large the noise will be and, in the case of corpus analysis, (c) how prevalent the phenomenon of interest is, and (d) whether the sampling technique planned will lead to bias (see also Rowland & Fletcher, 2006; Tamis-LeMonda et al., 2017; Tomasello & Stahl, 2004). We hope future methodological work on LFSE recordings can produce some rules of thumb for addressing these issues.

In budgeting for the actual time it takes to annotate, researchers should keep in mind that tasks differ in their complexity and difficulty. By and large, doing speaker segmentation and diarization can take between 4 and 20 times the recording time (i.e., 4–20 minutes to annotate 1 minute of audio), depending on the required temporal precision (how accurate should the onsets and offsets be); and the difficulty of the audio clip being annotated (when there is silence, it will go quickly; when there is lively conversation between four talkers, it will go slowly). Even tasks of the same type (e.g., multiple choice of who a sentence is spoken to, or whether the sentence is a declarative, question, or imperative) can vary enormously in how long it takes the annotator to make a decision and, ultimately, in how reliable annotations are. We therefore advise those embarking on annotation with limited resources (e.g., assistant time or funding) to carefully consider the cumulative annotation time necessary per minute of recording when deciding what to annotate, and in which portions of the full recording. In our experience, this estimation is best made by having annotators complete a few sample clips from variable subsets of the data.

6. Conclusions

We set out to comprehensively review the topics worth considering before embarking on a project centered on LFSE recordings. An emergent theme in our review was the critical need for more (and more consistent) shared annotated data. Without it, we cannot hope for better tools for automated analysis. For now, the answers to many research questions will still require significant manual annotation, and it is therefore paramount for researchers to plan ahead for long-term sharing long before data collection begins. Overcoming many of these challenges is likely best achieved through the effort of research teams and communities who can collectively address the larger theoretical and practical issues relevant to LFSE recording research. We hope that we have facilitated this process with our coverage of relevant issues above, but we may have missed some topics. Further, much of the information we have included is bound to become out-of-date as technology continues to develop. We encourage readers to join the DARCLE network (darcle.org) to stay in the loop about the latest developments in LFSE recording research. DARCLE is focused on child language, but many of the methodological considerations are universal to those considering LFSE recording studies, and exchanges between researchers working on different participant groups could be mutually beneficial. Readers can post about their own findings in that and similar mailing lists. A repository of papers using LFSE recordings has also been made available by Ganek (2018). Her paper repository is community-augmented, meaning that anyone can add their resources to the list when they are ready to share. We believe LFSE recording-based research is a fast-growing domain with enormous scientific potential. We hope that the next decade will see a continued shift toward open, shared tools and databases that can facilitate our understanding of everyday language environments.

Data Accessibility Statement

The papers and resources reviewed in this article can be accessed via the links provided.

Additional File

The additional file for this article can be found as follows:

- **Appendices.** A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. DOI: https://doi.org/10.1525/collabra.209.s1

Acknowledgements

We gratefully acknowledge our colleagues in ACLEW and DARCLE who have sparked the need for a paper along these lines and whose ideas and research has inspired us to take up LFSE recording projects of our own. We thank Meg Cychosz, Rachid Riad, Caroline Rowland, and Christof Neumann for helpful feedback on previous versions of this manuscript.

Funding Information

This work was supported by an NWO Veni Innovational Research Scheme grant (275-89-033) to MC and by a TransAtlantic Platform "Digging into Data" collaboration grant (ANR-16-DATA-0004 ACLEW: Analyzing Child Language Experiences Around The World) and Agence Nationale de la Recherche (ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017) and J. S. McDonnell Foundation grant to AC.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

Both MC and AC contributed at every stage of manuscript development, from conception to writing.

References

Abels, M., & **Vogt, P.** (2018). Speech acts addressed at Hadza infants in Tanzania. *Proceedings of the 12th International Conference on the Evolution of Language (Evolang12)*. DOI: https://doi.org/10.12775/3991-1.001

ACLEW/DiViMe. (2018). Retrieved from http://github.com/aclew/DiViMe.

ACLEW Project. (2018). Retrieved from https://sites.google.com/view/aclewdid/home.

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & **Tor, S.** (2019). Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science*, *22*(1), e12715. DOI: https://doi.org/10.1111/desc.12715

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. DOI: https://doi.org/10.1073/pnas.1712966114

Boersma, P. (2009). Praat: Doing phonetics by computer. Retrieved from http://www.praat.org/.

Bulgarelli, F., & **Bergelson, E.** (under review). Look who's talking: A comparison of automated and human-generated speaker tags in naturalistic daylong recordings.

Busch, T., Sangen, A., Vanpoucke, F., & **Wieringen, A. V.** (2017). Correlation and agreement between Language ENvironment Analysis (LENA™) and manual transcription for Dutch natural language recordings. *Behavior Research Methods*, *50*(5), 1921–1932. DOI: https://doi.org/10.3758/s13428-017-0960-0

Canault, M., Normand, M. L., Foudil, S., Loundon, N., & **Thai-Van, H.** (2016). Reliability of the Language ENvironment Analysis system (LENA™) in European French. *Behavior Research Methods*, *48*(3), 1109–1124. DOI: https://doi.org/10.3758/s13428-015-0634-8

Cartmill, E. A., Amstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & **Trueswell, J. C.** (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences 110*(28), 11278–11283. DOI: https://doi.org/10.1073/pnas.1309518110

Casillas, M., Bergelson, E., Warlaumont, A. S., Cristia, A., Soderstrom, M., Vandam, M., & **Sloetjes, H.** (2017). A New Workflow for Semi-Automatized Annotations: Tests with Long-Form Naturalistic Recordings of Children's Language Environments. In M. Wlodarczak (Ed.) *Proceedings of the 18th Annual Conference of the International*

Speech Communication Association (INTERSPEECH 2017), 2098–2102. DOI: https://doi.org/10.21437/Interspeech.2017-1418

Casillas, M., Brown, P., & **Levinson, S. C.** (2017). Casillas HomeBank Corpus. Retrieved from https://homebank.talkbank.org/access/Secure/Casillas.html.

Casillas, M., Brown, P., & **Levinson, S. C.** (under review). Early language experience in a Tseltal Mayan village.

Casillas, M., Bunce, J., Soderstrom, M., Rosemberg, C., Migdalek, M., Alam, F., Stein, A., & **Garrison, H.** (2018). Tutorials: Using the ACLEW DAS template. Retrieved from https://osf.io/b2jep/.

DARCLE. (2018). Retrieved from http://darcle.org/.

Databrary. (2018). Retrieved from https://www.databrary.org/resources/guide/investigators/authorization.html.

Datavyu Team. (2014). Datavyu: A Video Coding Tool. Databrary Project, New York University. Retrieved from http://datavyu.org.

Elo, H. (2016). Acquiring Language as a Twin: Twin children's early health, social environment and emerging language skills. PhD dissertation, Tampere University.

EU GDPR Information Portal. (2018). Retrieved from http://www.eugdpr.org/.

Fausey, C. M., & **Mendoza, J. K.** (2018). FauseyTrio-Public HomeBank Corpus. doi:10.21415/T56D7Q. Retrieved from https://homebank.talkbank.org/access/Public/FauseyTrio-Public.html.

Frank, M. C., Braginsky, M., Marchman, V. A., & **Yurovsky, D.** (2019). Variability and Consistency in Early Language Learning: The WordBank Project. Retrieved from https://langcog.github.io/wordbank-book/ May 2019.

Ganek, H. (2018). LENA Studies Spreadsheet. September 2018. DOI: https://doi.org/10.17605/OSF.IO/54FY7

Ganek, H., & **Eriks-Brophy, A.** (2018a). Language ENvironment analysis (LENA) system investigation of day long recordings in children: A literature review. *Journal of Communication Disorders*, *72*, 77–85. DOI: https://doi.org/10.1016/j.jcomdis.2017.12.005

Ganek, H. V., & **Eriks-Brophy, A.** (2018b). A Concise Protocol for the Validation of Language ENvironment Analysis (LENA) Conversational Turn Counts in Vietnamese. *Communication Disorders Quarterly*, *39*(2), 371–380. DOI: https://doi.org/10.1177/1525740117705094

Gilkerson, J., & **Richards, J. A.** (2008). LENA TR-02-2: The LENA natural language study. Boulder, CO: LENA Foundation.

Gilkerson, J., & **Richards, J. A.** (2009). LENA ITR-01-02: The power of talk: Impact of adult talk, conversational turns, and TV during the critical 0–4 years of child development. Boulder, CO: LENA Foundation. DOI: https://doi.org/10.1044/leader.IN1.14102009.4

Gilkerson, J., Zhang, Y., Xu, D., Richards, J. A., Xu, X., Jiang, F., Topping, K, et al. (2015). Evaluating Language Environment Analysis System Performance for Chinese: A Pilot Study in Shanghai. *Journal of Speech Language and Hearing Research*, *58*(2), 445. DOI: https://doi.org/10.1044/2015_JSLHR-L-14-0014

Git. (2018). Retrieved from http://git-scm.com/.

GitHub. (2018). Retrieved from https://github.com/.

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., & **Gilkerson, J.** (2010). Assessing Children's Home Language Environments Using Automatic Speech Recognition Technology. *Communication Disorders Quarterly*, *32*(2), 83–92. DOI: https://doi.org/10.1177/1525740110367826

Henrich, J., Heine, S. J., & **Norenzayan, A.** (2010). The weirdest people in the world. *Behavioral and Brain Sciences 33*(2–3), 61–83. DOI: https://doi.org/10.1017/S0140525X0999152X

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review 26*(1), 55–88. DOI: https://doi.org/10.1016/j.dr.2005.11.002

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics*, *3*(2), 119–131. DOI: https://doi.org/10.1007/s40708-016-0042-6

HomeBank. (2018). Retrieved from http://homebank.talkbank.org/.

HomeBankCode. (2018). Retrieved from http://github.com/homebankcode.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & **Hedges, L. V.** (2010). Sources of variability in children's language growth. *Cognitive Psychology 61*, 343–365. DOI: https://doi.org/10.1016/j.cogpsych.2010.08.002

Jackson, C. W., & **Callender, M. F.** (2014). Environmental Considerations. *Topics in Early Childhood Special Education*, *34*(3), 165–174. DOI: https://doi.org/10.1177/0271121414536623

KALDI for dummies. (2018). Retrieved from kaldi-asr.org/doc/kaldi_for_dummies.html. September 2018.

Karam, Z. N., Provost, E. M., Singh, S., Montgomery, J., Archer, C., Harrington, G., & **Mcinnis, M. G.** (2014). Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DOI: https://doi.org/10.1109/ICASSP.2014.6854525

Kisler, T., Reichel, U., & **Schiel, F.** (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347. DOI: https://doi.org/10.1016/j.csl.2017.01.005

Lee, Y., Song, J., Min, C., Hwang, C., Lee, J., Hwang, I., Lee, U, et al. (2013). SocioPhone. *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services – MobiSys '13*. DOI: https://doi.org/10.1145/2462456.2465702

Le Franc, A., Riebling, E., Karadayi, J., Wang, Y., Scaff, C., Metze, F., & **Cristia, A.** (2018). The ACLEW DiViMe: An easy-to-use diarization tool. In B. Yegnanarayana, C. Chandra Sekhar, S. Narayanan, S. Umesh, S. R. M. Prasanna, Hema A. P. Kumar Ghosh, et al. (Eds.)

Proceedings of Interspeech 2018 (pp. 1383–1397). DOI: https://doi.org/10.21437/Interspeech.2018-2324

Li, L., Vikani, A. R., Harris, G. C., & **Lin, F. R.** (2014). Feasibility Study to Quantify the Auditory and Social Environment of Older Adults Using a Digital Language Processor. *Otology & Neurotology*, *35*(8), 1301–1305. DOI: https://doi.org/10.1097/MAO.0000000000000489

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk (third edition): Volume I: Transcription format and programs, Volume II: The database. *Computational Linguistics*, *26*(4), 657–657. DOI: https://doi.org/10.1162/coli.2000.26.4.657

MacWhinney, B. (2007). The TalkBank Project. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora: Synchronic Databases, 1*, 163–180. Houndmills: Palgrave-Macmillan. DOI: https://doi.org/10.1057/9780230223936_7

Marchman, V. A., Martínez, L. Z., Hurtado, N., Grüter, T., & **Fernald, A.** (2016). Caregiver talk to young Spanish-English bilinguals: Comparing direct observation and parent-report measures of dual-language exposure. *Developmental Science*, *20*(1), e12425. DOI: https://doi.org/10.1111/desc.12425

Mastin, J. D., Ellwood-Lowe, M., Marchman, V., & **Fernald, A.** (2016). Quantity & quality of CDS at 18-months predicts later vocabulary and language processing. Talk presented at ICIS, New Orleans.

Mehl, M. R. (2017). The Electronically Activated Recorder (EAR). *Current Directions in Psychological Science*, *26*(2), 184–190. DOI: https://doi.org/10.1177/0963721416680611

Mehl, M. R., & **Pennebaker, J. W.** (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, *84*(4), 857–870. DOI: https://doi.org/10.1037/0022-3514.84.4.857

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & **Price, J. H.** (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, *33*(4), 517–523. DOI: https://doi.org/10.3758/BF03195410

Mehl, M. R., Vazire, S., Holleran, S. E., & **Clark, C. S.** (2010). Eavesdropping on Happiness. *Psychological Science*, *21*(4), 539–541. DOI: https://doi.org/10.1177/0956797610362675

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & **Pennebaker, J. W.** (2007). Are Women Really More Talkative Than Men? *Science*, *317*(5834), 82–82. DOI: https://doi.org/10.1126/science.1139940

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, *1*(1), 131–144. DOI: https://doi.org/10.1177/2515245917747656

Nicolao, M., Christiansen, H., Cunningham, S., Green, P., & **Hain, T.** (2016). A Framework for Collecting Realistic Recordings of Dysarthric Speech – thehomeServiceCorpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Nielsen, M., Haun, D., Kärtner, J., & **Legare, C. H.** (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38. DOI: https://doi.org/10.1016/j.jecp.2017.04.017

Norcliffe, E., Harris, A. C., & **Jaeger, T. F.** (2015). Cross-linguistic psycholinguistics and its critical role in theory development: Early beginnings and recent advances. *Language, Cognition and Neuroscience*, *30*(9), 1009–1032. DOI: https://doi.org/10.1080/23273798.2015.1080373

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Seligman, M. E., et al. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. DOI: https://doi.org/10.1037/pspp0000020

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Silovsky, J., et al. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding (No. EPFL-CONF-192584)*. IEEE Signal Processing Society.

Pye, C. (2017). *The Comparative Method of Language Acquisition Research*. The University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226481319.001.0001

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

Ramírez-Esparza, N., García-Sierra, A., & **Kuhl, P. K.** (2017). Look Who's Talking NOW! Parentese Speech, Social Context, and Language Development Across Time. *Frontiers in Psychology*, *8*: 1008. DOI: https://doi.org/10.3389/fpsyg.2017.01008

Ramírez-Esparza, N., Mehl, M. R., Álvarez-Bermúdez, J., & **Pennebaker, J. W.** (2009). Are Mexicans more or less sociable than Americans? Insights from a naturalistic observation study. *Journal of Research in Personality*, *43*(1), 1–7. DOI: https://doi.org/10.1016/j.jrp.2008.09.002

Robbins, M. L., Mehl, M. R., Holleran, S. E., & **Kasle, S.** (2011). Naturalistically observed sighing and depression in rheumatoid arthritis patients: A preliminary study. *Health Psychology*, *30*(1), 129–133. DOI: https://doi.org/10.1037/a0021558

Rodríguez-Arauz, G., Ramírez-Esparza, N., García-Sierra, A., Ikizer, A. G., & **Fernández-Gómez, M. J.** (2018). You go before me, please: Behavioral politeness and interdependent self as markers of simpatía in Latinas. *Cultural Diversity and Ethnic Minority Psychology*. Advance online publication. DOI: https://doi.org/10.1037/cdp0000232

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & **Gabrieli, J. D.** (2018). Beyond the 30-Million-Word Gap: Children's Conversational Exposure Is Associated With Language-Related Brain Function. *Psychological Science*, *29*(5), 700–710. DOI: https://doi.org/10.1177/0956797617742725

Rowland, C. F., & **Fletcher, S. L.** (2006). The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language 33*(4), 859–877. DOI: https://doi.org/10.1017/S0305000906007537

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., & **Liberman, M.** (2018) The First DIHARD Speech Diarization Challenge. In *Proceedings of Interspeech 2018*. Hyderabad, India.

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., & **Liberman, M.** (2019). The Second DIHARD Diarization Challenge: Dataset, task, and baselines. In *Proceedings of Interspeech 2019*. Graz, Austria.

Scaff, C., Stieglitz, J., Casillas, M., & **Cristia, A.** (in preparation) Language Input in a hunter-forager population: Estimations from daylong recordings.

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Zafeiriou, S., et al. (2017). The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring. In *Proceedings of Interspeech 2017*. DOI: https://doi.org/10.21437/Interspeech.2017-43

Schuller, B. W., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychosz, M., Schmitt, M., et al. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In *Proceedings of Interspeech 2019*. Graz, Austria.

Schwarz, I.-C., Botros, N., Lord, A., & **Marcusson, A.** (2017). The LENA™ system applied to Swedish: Reliability of the Adult Word Count estimate. In M. Wlodarczak (Ed.) *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, pp. 2088–2092. DOI: https://doi.org/10.21437/Interspeech.2017-43

Slobin, D. I. (2014). Before the beginning: The development of tools of the trade. *Journal of Child Language*, *41*(S1), 1–17. DOI: https://doi.org/10.1017/S0305000914000166

Sloetjes, H., & **Wittenburg, P.** (2008). Annotation by category – ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

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: https://doi.org/10.1371/journal.pone.0080646

Tamis-Lemonda, C. S., Kuchirko, Y., Luo, R., Escobar, K., & **Bornstein, M. H.** (2017). Power in methods: Language to infants in structured and naturalistic contexts. *Developmental Science*, *20*(6), e12456. DOI: https://doi.org/10.1111/desc.12456

Tomasello, M., & **Stahl, D.** (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language 31*(1), 101–121. DOI: https://doi.org/10.1017/S0305000903005944

van Alphen, P., Meester, M., & **Dirks, E.** (2017). LENA onder de loep. *VHZ Artikelen*, April 2017.

VanDam, M. (2018a). VanDam Public 5-minute HomeBank Corpus. Retrieved from https://homebank.talkbank.org/access/Public/VanDam-5minute.html. DOI: https://doi.org/10.21415/T5388S

VanDam, M. (2018b). VanDam Public Daylong HomeBank Corpus. Retrieved from https://homebank.talkbank.org/access/Public/VanDam-Daylong.html. DOI: https://doi.org/10.21415/T5QH5N

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

VanDam, M., & **Silbert, N. H.** (2016). Fidelity of Automatic Speech Processing for Adult and Child Talker Classifications. *Plos One*, *11*(8), e0160588. DOI: https://doi.org/10.1371/journal.pone.0160588

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., Palma, P. D., & **MacWhinney, B.** (2016). HomeBank: An Online Repository of Daylong Child-Centered Audio Recordings. *Seminars in Speech and Language*, *37*(02), 128–141. DOI: https://doi.org/10.1055/s-0036-1580745

VanDam, M., Warlaumont, A., MacWhinney, B., Soderstrom, M., & **Bergelson, E.** (2018). Vetting Manual: Preparation of Recordings for Unrestricted Publication in HomeBank (Version 1.1). DOI: https://doi.org/10.21415/T56H4M

Wang, Y., Neves, L., & **Metze, F.** (2016). Audio-based multimedia event detection using deep recurrent neural networks. In *The 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2742–2746). DOI: https://doi.org/10.1109/ICASSP.2016.7472176

Weisleder, A., & **Fernald, A.** (2013). Talking to Children Matters. *Psychological Science*, *24*(11), 2143–2152. DOI: https://doi.org/10.1177/0956797613488145

Wells, G. (1979). Describing Children's Linguistic Development at Home and at School. *British Educational Research Journal*, *5*(1), 75–98. DOI: https://doi.org/10.1080/0141192790050109

Winkelmann, R., Harrington, J., & **Jänsch, K.** (2017). EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language*, *45*, 392–410. DOI: https://doi.org/10.1016/j.csl.2017.01.002

Woynaroski, T., Oller, D. K., Keceli-Kaysili, B., Xu, D., Richards, J. A., Gilkerson, J., Yoder, P., et al. (2016). The stability and validity of automated vocal analysis in preverbal preschoolers with autism spectrum disorder.

Autism Research, *10*(3), 508–519. DOI: https://doi.org/10.1002/aur.1667

Xu, D., Yapanel, U., & **Gray, S.** (2009). LENA TR-05: Reliability of the LENA Language Environment Analysis System in young children's natural home environment. Boulder, CO: LENA Foundation.

Ziaei, A., Sangwan, A., Kaushik, L., & **Hansen, J. H.** (2015). Prof-Life-Log: Analysis and classification of activities in daily audio streams. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DOI: https://doi.org/10.1109/ICASSP.2015.7178866

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.209.pr

How to cite this article: Casillas, M., and Cristia, A. (2019). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra: Psychology,* 5(1): 24. DOI: https://doi.org/10.1525/collabra.209

Senior Editor: Rolf Zwaan

Editor: Mark Dingemanse

Submitted: 19 November 2018 **Accepted:** 30 April 2019 **Published:** 23 May 2019

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS