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

Marisa Casillas[a] and Alejandrina Cristia[b]

a. Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands,
b. Laboratoire de Sciences Cognitives et Psycholinguistique, Dept d'Etudes Cognitives, ENS, PSL University, EHESS, CNRS, Paris, France.
marisa.casillas@mpi.nl and alecristia@gmail.com

Appendix A: Further considerations for ethics and permissions in sensitive contexts

In this section we provide additional information about issues to consider before carrying out a study with a population that is in some way vulnerable or simply unfamiliar to the research team.

A.1. Ethical review and informed consent

All proposals for research with human participants must be rigorously assessed and approved through independent ethics institutions before data collection begins. In the US and Europe, institutional approval usually comes through an institutional review board or ethics committee: a collection of individuals who are responsible for reviewing research proposals from a particular research community, e.g., a university or research institute. The amount of oversight and regulation exerted by ethics institutions varies from place to place, but typically the goal is to ensure that any risk to participants is minimized and, when more than minimal, sufficiently justified and communicated to participants. Ethical committees need to assess this risk for the entire arc of each research project: data collection, annotation, analysis, long-term archiving, distribution of findings, and future intellectual property rights.

Those working with participants in vulnerable communities, such as traditional/indigenous, non-industrial, developing, and/or low-literacy communities, need to take extra care in communicating with their ethics institutions about planned research. To do their job, ethics committees must formally assess research risks. However, it is much easier for these committees (and indeed the researchers) to assess risk for populations and communities like their own, where the context of the research and participants' expectations are familiar to committee members. How can ethics committees effectively assess risk and advise on topics like consent and communication of participant rights when they are completely unfamiliar with the participant population? Take, for example, the case of written consent. Written consent is an excellent tool for ensuring that each participant gets the same information about the study and their rights (e.g., via an information sheet), and ends with a clear record of their informed consent (i.e., the signed sheet). However, in some populations (e.g., where literacy is low or legal-like formal procedures are uncommon), spoken, dialogic consent procedures are much more effective for protecting participant rights. In these cases, participants may instead be most effectively informed of their rights through an interactional exchange that allows them to ask questions in a way that feels comfortable and normal to them. Even in some literate communities, the insistence on written consent may raise significant suspicion among

Accepted on 30 April 2019 for publication in *Collabra*

participants, straining otherwise productive and positive relationships between the researchers and the community. We therefore highly recommend that researchers intending to conduct research in a new community get in touch with others who are experienced in fieldwork, especially if they have experience in the area/community of focus. Once researchers anticipate the needs for conducting an effective consent process in the community, they can argue for these processes when making their official ethics applications. One can find additional guidance in the ethical frameworks published by groups of researchers concerned with these issues, such as from the Linguistic Society of America (2009), the American Anthropological Society (2012; ethics forum at http://ethics.americananthro.org/about/), the DOBES group (Wittenburg, 2005), and many more.

A.2. Research visas for work abroad

Researchers collecting data abroad must also investigate whether they need to obtain a research visa. Research oversight varies dramatically from nation to nation. In some cases, there is no research visa process for behavioral social science, and so researchers can enter the country with a tourist visa. In other cases, nations may have extensive application processes in which the proposed research is scrutinized closely before granting any access and may. Such processes may even require direct collaboration with local institutions and/or sharing ownership of the resulting intellectual property. In developing countries, the process of obtaining a research visa can be unpredictable, and it is wise to set aside extra time and funding for applications to be processed. Researchers new to a community benefit enormously from talking to others who have worked in that same or a similar community (or even that same nation) about who to contact, what is required, and what to expect in the process of obtaining a visa. When possible, having local contacts to facilitate communication between researchers and visa-granting offices is ideal. Informally, some of the researchers we have talked to who work abroad have done their work with a mix of tourist and research visas, even when they should have only used research visas. In many cases, researchers were working on projects for which proper research visas were obtained at some time point for at least some of the collaborators. Using tourist visa when one should have a research or other visa carries an enormous amount of risk, including losing the ability to use the data, being barred from future entry into the community or nation, or other serious social and legal repercussions.

A.3. Additional permissions and considerations

Even if one obtains ethical approval from an institutional review board and a valid research visa, it may still be necessary to obtain formal or informal permission to do research in a community at the level of local government or leadership. For example, in some cases, researchers might first need to approach a council of community leaders, village elders, heads of households, etc., before seeking consent for participation from individuals within the community. Here again, prior knowledge is hugely beneficial to researchers new to a community in seeking out and obtaining permission from the appropriate local sources.

Researchers must also be wary of local power structures when obtaining individual consent and making decisions about how to hire research assistants and compensate participants and/or the community. We know of at least one case where permission from a village elder implies consent

for all community members, such that asking an individual for their consent can be seen as questioning their adherence to the village elder's commitment. In sum, even with permission at multiple levels, researchers must be vigilant in carrying out an ethical framework for individual participants, not only to achieve their own research goals ethically but to also establish an open and positive connection so that future researchers may also work with these communities.

A.4. Compensation at the individual and community level

Finally, throughout this process, researchers should carefully consider in what ways their presence could minimize inconvenience and, even better, benefit the community. Formally, this may include plans for participant and research assistant compensation, local dissemination of the results, and application of these results to community materials, such as books, dictionaries, and phone apps. Informally, this may include things like supporting community members who want to gain expertise in the research domain, bringing in and sharing expertise outside of the researcher's primary domain that may be useful (e.g., general information about dental hygiene or nutrition), and building strong interpersonal connections with individual community members.

Appendix B: Further information on archiving

In this section we provide additional information on methods for archiving, particularly with respect to version control and data curation.

B.1. Version control

Once a dataset has been collected and stored, it is likely to remain in a fairly stable state for the long term, but one should be ready to integrate changes in both the primary data and the annotations accompanying it. With respect to the primary data (i.e., the LFSE recordings themselves), participants may request for part of a recording to be silenced, or they may require the complete removal of a recording or a group of recordings. With respect to secondary data (e.g., annotations or metadata), further annotations or corrected annotations by the data holders or re-users is likely to require updates to the stored data. In the case of LFSE recordings, one obvious way in which annotations are enriched is in terms of temporal coverage: at the beginning of a project, it may be the case that less than 1% of the recording is annotated, but as the project progresses, more annotation coverage may grow substantially. Annotations often get enriched "vertically" as well: at later stages of a project researchers might add information about speech environment, interlocutors, discussion topic, syntactic information, and more to an annotation file that originally only included an initial orthographic transcription. In truth, there is no such thing as a "final" version of an annotation.

In the context of cumulative and replicable science, it would be ideal for each version to be independently citable, so that someone seeking to replicate a result can go back in time and re-run analysis on the data as they were at a given point in time. Additionally, re-users may want to be told of new versions of the annotations, so that they can expand their analyses or make sure they run them on the most recently corrected data. It would be ideal if re-users were able to give back in terms of re-annotation or correction. One obvious example of this pertains the task of "speaker diarization", determining where in a long file someone is talking. At the time of writing,

there is still no open source and perfectly accurate method for doing speaker diarization on daylong recordings. Thus this task is still best performed by human annotators. But as daylong recording repositories become available, speech technologists may be able to develop algorithms for automatically providing such annotations by re-using existing data. It would be ideal if speech technology re-users were able to return to the original data holders their new automatic annotations, whose performance can be established against the human-annotated portion of the data set. In this scenario, everyone benefits: the original data holders and other data users get new annotations with which they can choose to work (if the accuracy is sufficient for their purposes), and the speech technologists make advances in their own research domain.

To our knowledge, no scientific data repository allows the level of version control that we have set out here as the ideal, but we are hopeful that one will be developed in the near future. In fact, the technology enabling this functionality has existed for some time, thanks to Git (Git, 2018), which is both free and open source. Git allows version control, flagging of issues, following of repositories (with update and issue alerts), and even submission of improvements via a mechanism called a "pull request". There are many Git tutorials available on the web (we recommend, e.g., one created by the Software Carpentry: https://software-carpentry.org/, a non-profit aiming to empower researchers with programming skills). For those who are uncomfortable with using the command line, it is also now possible to use Git with GitHub (an online repository system that uses Git) through a graphical user interface that looks like a typical user-friendly software application (https://desktop.github.com/; GitHub, 2018). Finally, specific recommendations regarding how the idealized system could work have been laid out years ago (https://blog.okfn.org/2013/07/02/git-and-github-for-data/), and have been implemented on the Data Hub (Data Hub, 2018) for smaller, non-audio datasets. By combining archives that can accommodate raw audio data (as discussed in the main text) with the logic behind Data Hub, we could have an archive of audio files and also continuously improve and version annotations.

B.2. Data curation

All of the systems described above and in the main text (i.e., OSF, Databrary, and HomeBank), including those built on Git, rely on central curation: one or a few individuals check and vouch for data quality. In the case of OSF and Databrary, there is no external check and thus data holders are free to implement whatever annotation system they think is best. In the case of HomeBank, at the present time there are a small number of individuals who are well-versed in the CHAT format and may signal to data contributors and/or correct for data contributors' gross deviations (e.g., that a transcription noted as being orthographic is in fact phonetic). But even they may not notice other violations of data quality (e.g., inaccurate time stamping). In a git-based system, re-users can at least log these problems and propose solutions. However, *someone* in charge has to take these into account, and make an informed decision about implementing changes. In the short term, data contributors and database managers are the natural group to take the lead on these decisions. But how can we ensure the longevity of these datasets, 50, 100, or more years down the line? Eventually, it may be necessary to adopt a distributed curation system, like that implemented in wiki systems, where users can signal issues, vote for solutions, and more.

Appendix C: Further information on LENA accuracy

In this section we give greater detail on the LENA system's reported accuracy levels with respect to talker diarization (i.e., who talks when), evaluations of its accuracy in populations different from the one originally trained on, and limits of these evaluation measures worth considering in future work.

C.1. Talker diarization

In evaluating automated annotations, speech technologists often report two metrics in particular: *recall* (also called *sensitivity)* and *precision* (*specificity*). For example, imagine we are evaluating the cases where LENA's system marked a clip as speech from the "Target child"; *recall* is the percentage of clips that the human coder classified as being the target child and that the LENA system also classified as the target child. In other words, it measures the proportion of clips containing target child speech that the LENA system successfully found. *Precision* is the percentage of clips that the LENA system identified as the target child that were indeed spoken by the target child and not by someone else. There is a tradeoff between these two: one can achieve 100% recall and capture all the times the child spoke by simply classifying the whole of the daylong recording as being "Target child"; but in this case the precision would be very low, since the clips labeled as "Target child" would then include many silences and instances of speech from other speakers.

In the context of deciding whether LENA software is good enough for one's research purposes, researchers should consider how important recall and precision are for the planned annotation and analyses. For instance, imagine that I find that, in my sample, for regions that have been automatically tagged as being the target speaker, precision is .7 and recall is .4. This means that, among the clips automatically tagged as being the target speaker, 70% were really from that person; and that the automatic system found 40% of all the clips in which the target speaker was speaking. If my goal is to use the automatic counts to extract high-volubility sections that I will subsequently hand-code for syntactic complexity, then this level of performance is perfectly acceptable. Indeed, when hand-transcribing, I can easily ignore the 30% "garbage" that was incorrectly tagged. However, if my goal is to estimate how much children speak in this community, or to study individual variation in the community, then I should check whether the .4 recall is stable across the children in the sample, in which case I can simply apply a conversion to make up for the vocalizations that are missed by the algorithm. If I find that the .4 arises from some children being detected at 90% and others at 10%, it implies that the conversion factor may be fine at the population level but it is not suitable to interpret speech patterns at the individual level.

Table C1 shows LENA's precision and recall performance in seven separate reports on child speech environments (Bulgarelli & Bergelson, under review; Elo, 2016; Gilkerson et al., 2015; Ko, Seidl, Cristia, Reimchen, & Soderstrom, 2016; Xu, Yapanel, & Gray, 2009; VanDam & Silbert, 2016; Seidl et al., 2018). Note that all of these studies sampled from American English learners, with the exception of Elo (2016), who used data from Finnish-acquiring children, and Gilkerson and colleagues (2015), who used data from Mandarin Chinese-acquiring children.

Notice that while the overall recall is high throughout, the precision varies to a greater extent, and is clearly lower for the two non-American English samples, reaching very low levels for the characterization of adults versus children in the Mandarin dataset. Thus, accuracy is affected by a mismatch between the population the LENA software was developed for and the population with which it was used—even with a task that seems at first glance to be somewhat language independent.

Table C1: Recall and precision of the talker label attributed to a stretch of speech. Columns represent the different studies (Bul: Bulgarelli & Bergelson, under review (Am. English); Elo, 2016 (Finnish), Gilk: Gilkerson et al., 2015 (Mandarin), Ko: Ko et al., 2016 (Am. English), LTR: LENA Technical Report #5 (Xu et al., 2009; Am. English), VD: VanDam & Silbert, 2016 (Am. English), Seidl: Seidl et al., 2018 (Am. English). In the table, (c) indicates that the two relevant rows have been collapsed in the evaluation.

	Recall			Precision					
	LTR5	Gilk	Elo	VD	Ko	Bul	Seidl	Elo	Gilk
Target child (Child)	76% (c)	79% (c)	90%	86% (c)	88% (c)	59% (c)	72%	58%	21% (c)
Any other child (OCh)			86%				NA	94%	
Any female adult (FA)	82% (c)	81% (c)	83%	60%	83%	73% (c)	72%	95%	66% (c)
Any male adult (MA)			91%	60%			NA	96%	

C.2. LENA evaluations for non-normed populations

One strength of the LENA literature is that numerous studies have evaluated the accuracy of the LENA system's output in populations other than the normative American English sample (Xu et al., 2009). Overall, these independent evaluations have shown very good results for adult word count with correlations reported for American English being above .8, and for other languages ranging from .6 to .9 (Weisleder & Fernald, 2013 for Spanish; Canault, Normand, Foudil, Loundon, & Thai-Van , 2016 for French; Schwarz, Botros, Lord, & Marcusson, 2017 for

Swedish; Gilkerson et al., 2015 for Mandarin; Busch, Sangen, Vanpoucke, & Wieringen, 2017 for Dutch; Ganek & Eriks-Brophy, 2018 for Vietnamese; see also reported error estimates in Elo, 2016 for Finnish; Gilkerson et al., 2015 for Mandarin; van Alphen, Meester, & Dirks, 2017 and Busch et al., 2017 for Dutch).

There are fewer evaluations of the other two global measures: number of vocalizations by the child (Child Vocalization Counts) and number of turns between the target child and an adult (Conversational Turn Counts). Initial results suggest that accuracy is more variable on these measures, with correlations between .65 and .75 for American English samples (Soderstrom & Wittebolle, 2013, Xu et al., 2009) and a great deal more variable for non-American English samples (e.g., Canault et al., 2016 for French; Gilkerson et al., 2015 for Mandarin; Busch et al., 2017 for Dutch).

C.3 Limits on the usefulness of these derived metrics

Focusing solely on precision and recall for individual tasks is not sufficient for evaluating error in audio processing pipelines that conduct multiple tasks. Imagine, for instance, what happens when a vocalization that was tagged as an adult actually came from a child. The diarization is in error on its own, but, as a consequence, the utterance will then be given an adult word count by LENA's system, passing the error on to a different stage in the processing pipeline—because it was child speech, it should not have been given a count at all. It is unclear how to best evaluate errors in speech processing pipelines because they can have these kinds of cascading consequences. Unlike speech technology areas focused on single tasks, such as voice activity detection and speaker diarization, there are no standards as to how to evaluate interconnected, multi-stage annotation workflows, and it thus remains an area much in need of further methodological research.

A separate issue is how clips of audio are selected for evaluation. In the past, most evaluators have purposefully avoided regions of the recordings they found to be noisy. For instance, evaluators have extracted 5-minute chunks that the LENA has labeled as having high adult word counts. Since this constitutes selective sampling, it produces estimates of accuracy that are unlikely to generalize to the recording as a whole.

Appendix D. Further information on DiViMe

In this section we give further information on DiViMe, a developing open-source alternative to the LENA automated annotation system. In 2018, the DiViMe team (Le Franc et al., 2018; ACLEW/DiViMe, 2018) published a paper describing their system, benchmarking it against LENA, and against state-of-the-art solutions using a metric called Diarization Error Rate (DER). DER is a standard metric proposed by the National Institute of Standards and Technology (Fiscus, Ajot, Michel, & Garofolo, 2006), which evaluates both the accuracy of the voice activity detection and the classification of voiced stretches into the right talker labels (e.g., "Anne" versus "Robert"). To calculate DER, the audio is first split into 100 ms frames, each of which is evaluated for accuracy. As a result, DER can go from 0% to 300%, representing the sum of the **miss rate** (percent 0–100% of frames that contained speech, but that the system labeled them

as non-speech or silence), the **false alarm rate** (percent 0–100% of frames that were labeled as speech, but that in truth were not, e.g., silence labeled as target child speech), and the **incorrect attribution rate** (percent 0–100% of frames labeled as one speaker, that were, in fact, another speaker, e.g., speech by a child labeled as adult speech).

A LENA-DiViMe comparison was carried out on a dataset composed mainly of clips extracted from LFSE recordings using periodic sampling. As noted in the main text, since periodic sampling is independent from the speaker diarization system being used, estimates from this are unbiased and are more likely to generalize to unseen data. On this dataset, LeFranc and colleagues (2018) report a DER between 90% and 143% for the LENA system, and 72% to 151% for the systems included in DiViMe at the time. Thus, performance for LENA and DiViMe was quite comparable.

The same paper discusses the performance of DiViMe's tools in the "DIHARD Challenge", a competition where speech technology teams compete for the best performance in diarizing audio data, and which included a range of data (including business meetings as well as random extracts from LFSE recordings; Ryant et al., 2018). Since this test set contained a mixture of several corpora, most of which were not collected with LENA, the LENA algorithms could not be applied, but the DiViMe algorithms could. The DER for DiViMe on the DIHARD challenge test data was 72%. In contrast, the best systems submitted to the competition scored 24% (DiHard leaderboard; lower is better). This difference–72% vs. 24% diarization error rate—is a first indication of how much LFSE recording researchers stand to gain by adopting state-of-the-art systems.

Currently, the developers of DiViMe are working with the two top teams from the DIHARD 2018 competition towards incorporating their winning systems, as well as further alternatives for each step in the processing pipeline. Since August 2018, DiViMe contains four alternative voice activity detection routines (including Hansen and colleagues' TO Combo SAD, Sadjadi & Hansen, 2013; Ziaei, Sangwan, Kaushik, & Hansen, 2015), and a tool to categorize audio stretches into 17 sound categories (including speech, music, singing, and many others such "noisemes"; Wang, Neves, & Metze, 2016). Further progress is expected as the DIHARD 2019 challenge continues to include LFSE extracts (Ryant et al., 2019), and the speech technology community continues to express an interest in working with these challenging data (see also Schuller et al., 2017; 2019).

Appendix E: Recommendations for database management

In this section, we give a few tips for file naming, particularly with respect to the researcher's need to document the source of short clips and the version of annotations made.

E.1. Labeling whole recordings and subclips

There are a number of basic practices that should be used by any researcher not using a database management system. For example, a good file name will have a code for the participant and, if it has a date, the date will be in ISO, 8-digit format (YYYYMMDD, e.g.,

20180406, which is unambiguous, unlike 04062018, which could be in April or in June). Often, researchers using LFSE recordings do not annotate the whole daylong recording, but instead extract clips for annotation. There are a number of free audio editing products that would allow them to do extract clips efficiently, such as ffmpeg or Praat (Boersma, 2009), both of which can be used with a user-friendly graphical user interface (GUI) or a script, as preferred by the user. If the input is audiovisual, ffmpeg is probably the best choice.

While it makes sense to annotate extracted clips rather than loading the whole recording file, it also makes sense to implement a system where one can import their sparse annotations into a file that has the same timing structure as the original, whole recording file. Thus, we recommend that the user reflects on the best clip naming strategy in advance, so as not to find themselves with a set of files like "c01_clip3", where one does not know where that clip starts within the long recording. During initial clip extraction, it is usually easy to name it with the number of seconds or milliseconds that have elapsed from the recording onset to the onset of the clip. For example, a clip starting at 2hr 35min into the recording could get a start time of 09300—the label uses five digits because there are 86400 seconds in a 24-hour period and files named in this way are guaranteed to sort correctly if the same digit span is used for all of them.

E.2. Documenting versions

As mentioned in the main text, one additional, major issue is how to deal with the fact that annotations are never final. Further uses of the data will result in new and/or edited annotations. For this issue, our best recommendation is to learn how to use Git, which is a useful tool for all aspects of project development (LFSE recording annotation, but also for data analysis, paper-writing, and more). There are many tutorials for learning Git that are publicly available. We particularly recommend a very short and easy-to-follow from the Software Carpentry group (https://swcarpentry.github.io/git-novice/).

Appendix F: Annotation software overview

In this section we provide additional information for researchers to consider in selecting annotation software for their LFSE-recording-based project.

Choice of annotation software will depend on a few things (see Table F1 for options). The first is how much time and effort researchers/their annotators can invest in learning a new piece of software. If the user does not have previous experience scripting and has little time, they may be best served by software that has "easy" modes based on a user-friendly graphical user interface (GUI). Others may find some time or work with a collaborator who scripts, in order to make use of more advanced software capabilities.

The second key question to ask oneself is: What is the ideal target of coding? Researchers' analytical goals helpfully constrain the range of software options. Some researchers are interested in acoustic-phonetic features of the speech, and thus will have specific requirements, such as the ability to see a waveform and/or spectrogram. If instead the user is interested in

categorizing, e.g., noting negative versus positive emotion, it may be desirable to have a system that requires users to select annotation values from a closed set of options.

Researchers interested in linguistic structures often want to study relationships across different levels of representation, for instance, in phonetics, one may want to tag both words and the phones within them. In syntactic analyses, one may want to represent both sentences and the words within them. In conversational dynamics, one may want to separately represent the turns by one and another talker. All of these goals require multi-level representation, sometimes requiring a hierarchical relationship between levels (e.g., phones occur within words, which occur within turns). A number of pieces of software allow for multi-level annotation. Many require annotations in each tier to be timed, allowing overlap and other temporal analyses, but also requiring the user to make decisions on when the annotated behavior starts and stops. Only a few pieces of software represent hierarchies explicitly (see Table F1).

Finally, the user may require more than one type of annotation software. In this case, it may be worthwhile exploring the interoperability profile of the relevant software options. If two pieces of software are “interoperable”, it means that users can export from one format to another, or import a file from one application into another application. We strongly recommend users to extensively test such conversion tools because, as software updates come in, conversions may not always stay up-to-date, and the result may be faulty imports/exports. Most of the annotation programs we mention in Table F1 have helpful development teams, who in the past have demonstrated willingness to work with users to improve successful importing.

Table F1. Free, multi-platform (e.g., Windows/Mac/Linux) applications for annotation.
System: *support = currently being supported; OS = open source*
Input: *A=audio, AV=audiovisual, AV+=audiovisual and other kinds of time-stamped data*
Multitier: *Whether tiers are necessarily time-stamped, cannot be time-stamped, or both (i.e., user decides which tiers are time-stamped and which are not)*
Closed vocab: *Whether one can specify a set of categories and force the annotator to use one of them ("multiple choice")*
Free text: *Whether free text entry (e.g., transcription, comments) is possible*
Spectrogram: *Whether the coder can view an audio spectrogram while coding (useful for phonetic measurements)*
Large files: *Whether large files (e.g., > 0.5GB) can be easily opened and annotated*
Interoperability: *Other software from which annotations read in or exported out*
Modes: *Whether the software has a basic mode only, or it also offers advanced options (e.g., scripting)*

	Key strength	System	Input	Multi-tier	Closed vocab	Free text	Spectro-gram	Large files	Inter-operability	Modes
Praat (Boersma, 2009)	ideal for acoustic phonetics	support	A	timed	no	yes	yes	limited	CLAN, Phon	both
Phon (Rose et al. 2007)	ideal for phonological level	support, OS	AV	both	yes	yes	no	yes	Praat, CLAN	both
TranscriberAG	"recommended" by LENA	OS	A	timed	no	yes	yes	yes	none	easy
Datavyu (Datavyu Team, 2014)	User-defined key strokes	support, OS	AV+	untimed	yes	no	no	no	none	both
ELAN (Sloetjes & Wittenburg, 2008)	Multi-stream, use of template, interoperable	support, OS	AV	both	yes	yes	no	yes	CLAN, Praat Transcriber AG, ...	both
CLAN (MacWhinney, 2000)	Ideal for lexicon and grammar	support	AV	untimed	no	yes	no	yes	Praat, Phon, ELAN	both

It is important to note that, because interoperability is sometimes possible, an initial use of one application may not close the doors to all others. For example, the DARCLE Annotation Scheme (Casillas, Bergelson et al., 2017) was designed within ELAN but is planned for compatibility with CLAN via the use of CHAT-formatted ELAN templates (MacWhinney, 2000). This means that

researchers can take advantage both of the ELAN features (e.g., use of templates) and, eventually, CLAN features (e.g., automated MLU calculations).

Appendix G: How much data should one annotate when validating aggregate data?

In this section we point toward prior work to help LFSE researchers estimate the quantity of manually annotated data needed to validate estimates over large portions of their recordings. In answering this question, there are many approaches. And, given that this method is still relatively in its infancy, there is no approach that works equally well for all phenomena of study in all recording contexts. A broad rule of thumb may be to annotate samples of a similar size to what is used in previous work looking at similar phenomena. For example, looking through five papers evaluating mainly adult word count (AWC) estimates, we found the modal number of minutes per child is 60 (typically 12 segments, each 5 minutes long), with 4 to 22 children contributing recordings (Table G1). So those interested in doing validation for AWC estimates can base their own design on this prior work.

*Table G1: Clip sampling used to assess LENA performance. Duration is given in minutes. *Busch et al. extracted 2–5 clips from 6 recordings, with 1 recording from each child except for one child who contributed 2 recordings (total 48 clips).*

Reference	Data annotated for LENA validation
Weisleder & Fernald (2013)	12 clips x 5 min x 1 recording x 10 children = 600 min
Canault et al. (2016)	6 clips x 10 min x 3 recording x 18 children = 3240 min
Gilkerson et al. (2015)	1 clips x 15 min x 1 recording x 22 children = 330 min
Busch et al. (2017)	~2 clips x 5 min x 1 recording x 5 children* = 240 min
Schwarz et al. (2017)	12 clips x 5 min x 1 recording x 4 children = 240 min

Appendix H: Example studies

In this section we briefly describe four fictional scenarios in which a researcher is considering using LFSE recordings and is using the key decisions flowchart (main text Figure 1) to decide how best to design their research program. These examples are meant to get interested researchers started on their own use of the flowchart.

Scenario A: I am a researcher interested in using LFSE recordings to measure how often school-aged children encounter acoustic noise that may be damaging to their developing auditory systems. Because I am interested in recordings made at-home, at-school, and in other typical contexts in which children find themselves, my review of available corpora on HomeBank has turned up no relevant leads for existing LENA recordings. I join the DARCLE network to ask whether anyone has data of these kind that they would be willing to share and, indeed, I find

someone who has collected the kind of recordings I am looking for and who is willing to collaborate on my project. Because I am studying the acoustic profile of the children's environments, however, I cannot use the output provided directly by LENA's system. Instead I will have to take measurements through some other means. I do not have a large amount of funding and I am not myself technically savvy, but I do have access to masters students in speech language pathology who need to complete internships and who can be taught to extract and measure short clips of audio in a systematic way. Since the acoustic environment can be measured at any point in the recording and because I am interested in both global characterizations and measures, I estimate that I will have enough data to reliably measure acoustic noise in the LFSE recordings if I take a measurement once every 10 minutes for each recording. I am working with the original data holder to ensure that these annotations, though they are not useful for her immediate research questions, are stored with the other annotations relevant to the audio recordings. Thankfully, I found out that she is still collecting data, so I plan to budget for one visit to her recording site to make validation measures of some typically recorded acoustic environments using standard equipment from my field.

Scenario B: I am a researcher interested in adolescent emotional development. I am conducting a longitudinal study on adolescents at a high school nearby my university. I would like to add LFSE recordings to my research design, but have been unable to attain ethical approval that satisfies both my university's IRB and the administrators at the high school. Even if I were able to get approval, though, I realized that my topic of interest would require many hours of manual annotation that I cannot personally complete and that I have no way to pay for, as my current research funds are quite limited. For that reason I decided instead to use a simple periodic self-report system in which, with parental permission, participants carry a small buzzer with them throughout the day and log their activity and emotional state in a small, portable journal every 20 minutes. This has been enough to supplement my other analyses and has been both convenient and cheap!

Scenario C: I am a researcher interested in using LFSE recordings to measure loneliness in rural- and urban-dwelling adults. I have comprehensively checked available corpora and even asked DARCLE members if there are relevant existing corpora, but I have come across nothing suitable for my needs, LENA or otherwise, because I am interested in pairing my recordings with a questionnaire I have developed to measure loneliness. After many iterations of application with my local ethics committee, including special consideration of lonely participants as a possible vulnerable population, I have finally acquired permission to record participants. Because I plan to collect LFSE recordings from 50 participants between ages 25 and 50 in each sample (urban and rural; nearly 800 hours of audio) and because my analyses depend on annotations of (a) the individuals talking to the participant and (b) the topics of conversation, I have decided to use an alternative (i.e., non-LENA) system. Thankfully, I have a large grant to cover the costs of manual annotation, which I believe will be around $180k USD based on a sample pilot annotation session using the coding scheme which I have used several times in the past, and focusing on four times of day of interest: early morning, midday, mid-afternoon, and late evening. I have thought long and hard about the analyses I'd like to do to both describe who the adults talk to, how often, and about what (and how those three measures relate to the

loneliness questionnaire). I have worked on urban populations in the past, so can anticipate what some of these data will look like (how often talk of different type occurs, how variable questionnaire responses are, etc.). However, I am not sure how these estimates will generalize to the rural population. I have therefore based my 50-participant plan and four-time-a-day sampling scheme on a conservative estimate of 200% of my anticipated power for urban populations.

Scenario D: I am a researcher working on the effectiveness of hearing aids in the different acoustic environments faced by users in everyday life. I am interested in the properties of specific hearing aid devices and so was unable to locate an existing corpus for re-use. I have technical support at my institution, plus available research assistant time and other financial resources. Because I don't plan to make use of any of LENA's automated output, I decided to go with a custom system in which the participant wears a microphone that records short periods of audio every few minutes. I wasn't able to get ethical approval to keep the recordings given the sensitive nature of my participant sample. However, with some advice from my institution's technical support team, I have figured out how I can transmit the short clips in real time to a trained annotator who then classifies the auditory scene in real time using a (previously) validated scheme before deleting the clip altogether. This method is still expensive and complex, but it is ethical, suitable to my research needs, and makes concerns about future data storage much simpler.

References

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

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

American Anthropological Society (2012). Principles of Professional Responsibility. Retrieved from http://ethics.americananthro.org/category/statement/

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: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:10.3758/s13428-015-0634-8

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)*, p. 2098–2102. doi:10.21437/interspeech.2017-1418

Data Hub (2018). Retrieved from https://datahub.io/

Accepted on 30 April 2019 for publication in *Collabra*

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.

Fiscus, J. G., Ajot, J., Michel, M., Garofolo, J. S. (2006) The Rich Transcription 2006 Spring Meeting Recognition Evaluation. In Renals S., Bengio S., Fiscus J. G. (Eds.) *Machine Learning for Multimodal Interaction. MLMI 2006. Lecture Notes in Computer Science, vol 4299*. Springer, Berlin, Heidelberg. doi: 10.1007/11965152_28

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

Gilkerson, J., Zhang, Y., Xu, D., Richards, J. A., Xu, X., Jiang, F., . . . Topping, K. (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:10.1044/2015_jslhr-l-14-0014

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

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

Ko, E., Seidl, A., Cristia, A., Reimchen, M., & Soderstrom, M. (2016). Entrainment of prosody in the interaction of mothers with their young children – ERRATUM. *Journal of Child Language, 43*(04), 964–965. doi:10.1017/s0305000915000410

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 (Eds.) *Proceedings of Interspeech 2018* (pp. 1383–1397). doi:10.21437/Interspeech.2018-2324

Linguistic Society of America (2009). Ethics Statement. Retrieved from http://www.linguisticsociety.org/sites/default/files/Ethics_Statement.pdf

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:10.1162/coli.2000.26.4.657

Rose, Y., Hedlund, G. J., Byrne, R., Wareham, T., & MacWhinney, B. (2007). Phon 1.2. *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition - CACLA '07*. doi:10.3115/1629795.1629798

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.

Sadjadi, S. O., & Hansen, J. H. L. (2013). Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux. *IEEE Signal Processing Letters, 20*(3), 197–200. doi: 10.1109/LSP.2013.2237903

Schuller, B. W. Batliner, A. Bergler, C., Pokorny, F. B., Krajewski, J., Cychosz, M., . . . Schmitt, M. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian

Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. *In Proceedings of Interspeech 2019*. Graz, Austria.

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

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: 10.21437/interspeech.2017-1287

Seidl, A., Cristia, A., Soderstrom, M., Ko, E. S., Abel, E. A., Kellerman, A., & Schwichtenberg, A. J. (2018). Infant–Mother Acoustic–Prosodic Alignment and Developmental Risk. *Journal of Speech, Language, and Hearing Research, 61*(6), 1369–1380.

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:10.1371/journal.pone.0080646

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

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:10.1109/ICASSP.2016.7472176

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

Wittenburg, P. (2005). Code of Conduct: DOBES-CoC-V2. Retrieved from http://dobes.mpi.nl/ethical_legal_aspects/DOBES-coc-v2.pdf

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:10.1109/icassp.2015.7178866