A New Workflow for Semi-automatized Annotations: Tests with Long-Form Naturalistic Recordings of Childrens Language Environments

Marisa Casillas[1], Elika Bergelson[2], Anne S. Warlaumont[3], Alejandrina Cristia[4], Melanie Soderstrom[5], Mark VanDam[6], and Han Sloetjes[1]

[1]Max Planck Institute for Psycholinguistics, The Netherlands
[2]Duke University, USA
[3]Cognitive and Information Sciences, University of California Merced
[4]Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, Paris, France.
[5]University of Manitoba, Canada
[6]Washington State University and Hearing Oral Program of Excellence (HOPE) of Spokane, USA

marisa.casillas@mpi.nl, elika.bergelson@duke.edu, alejandrina.cristia@ens.fr

Abstract

Interoperable annotation formats are fundamental to the utility, expansion, and sustainability of collective data repositories. In language development research, shared annotation schemes have been critical to facilitating the transition from raw acoustic data to searchable, structured corpora. Current schemes typically require comprehensive and manual annotation of utterance boundaries and orthographic speech content, with an additional, optional range of tags of interest. These schemes have been enormously successful for datasets on the scale of dozens of recording hours but are untenable for long-format recording corpora, which routinely contain hundreds to thousands of audio hours. Long-format corpora would benefit greatly from (semi-)automated analyses, both on the earliest steps of annotation—voice activity detection, utterance segmentation, and speaker diarization—as well as later steps—e.g., classification-based codes such as child-vs-adult-directed speech, and speech recognition to produce phonetic/orthographic representations. We present an annotation workflow specifically designed for long-format corpora which can be tailored by individual researchers and which interfaces with the current dominant scheme for short-format recordings. The workflow allows semi-automated annotation and analyses at higher linguistic levels. We give one example of how the workflow has been successfully implemented in a large cross-database project.

Index Terms: daylong recordings, language acquisition, annotation, speech recognition, speaker diarization

1. Introduction

Thanks to CHILDES [1], established in 1984, developmental language scientists have been able to inspect each others' recordings and re-use hard-to-collect data for decades. A crucial factor in the success of CHILDES has been getting researchers to use a unified transcription scheme—namely CHAT (Codes for the Human Analysis of Transcripts)—which, like other unified schemes, is instrumental for effectively producing and analyzing data within a shared framework. CHAT's success is partly due to its flexibility regarding the type and detail of annotation required, and partly due to its companion tool CLAN (Computerized Language ANalysis), which facilitates searches and analyses over completed transcripts. Indeed, a few lines of CLAN run over a set of CHAT transcripts can perform dictionary look-ups for morphological annotation, part of speech tagging, and initial syntactic analyses, not to mention a host of data extraction techniques for more detailed analyses at these and other levels.

The CHILDES system evolved around collections of relatively short recordings, on the order of a few dozen hours, which could feasibly be fully manually annotated for each speaker's utterances (onset/offset, content, and supplementary information). But recent years have seen an explosion in long-form ("daylong") recordings, which aim to capture the full gamut of linguistic input in a child's "typical day". Long-form recordings have become easy to gather, but present a new set of analysis challenges. For example, they tend to contain long silences or irrelevant noise, involve multiple (often overlapping) speakers and quickly accumulate to hundreds or even thousands of hours of audio. Traditional exhaustive transcription workflows such as CHAT are therefore ill-suited for long-form recordings and, consequently, current analysis tools for CHAT transcripts (CLAN) are not readily applicable. LENA©, a popular system for long-format recordings, instead takes a big-data approach, running algorithms over recordings to estimate their contents without requiring any manual annotation. Unfortunately, LENA algorithms are not open source and cannot be adapted to new corpora and languages. Neither can they incorporate newer, state-of-the-art solutions. Thus there is no current approach that flexibly meets researchers' needs for long-format recordings.

Together with colleagues in the Daylong Audio Recordings of Children's Language Environment (DARCLE) network (see `darcle.org`), we present the DARCLE Annotation Scheme (DAS), a workflow specifically designed for annotating long-form natural language environment recordings (though equally usable with traditional short-form recordings), with subsequent algorithmic analyses in mind. DAS is made to be flexible such that it can be tailored to the goals and means of individual researchers yet maintain a core infrastructure and documentation repository to facilitate sharing, large-scale data analysis, and tool development. DAS is CHAT-compatible, a key feature for interoperability with current data sharing for short-format recordings in developmental language science. In what follows, we describe our workflow, illustrate how it can be adapted and implemented, and discuss its advantages, limitations, and future directions.

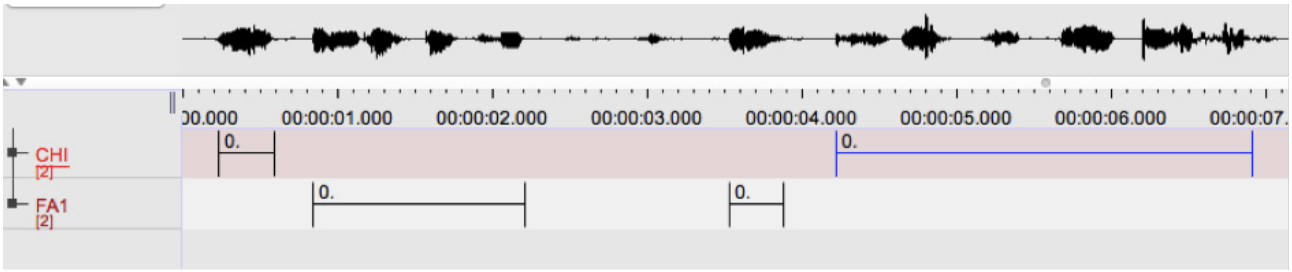

Figure 1: *Example coded audio segment with minimal DAS annotation.*

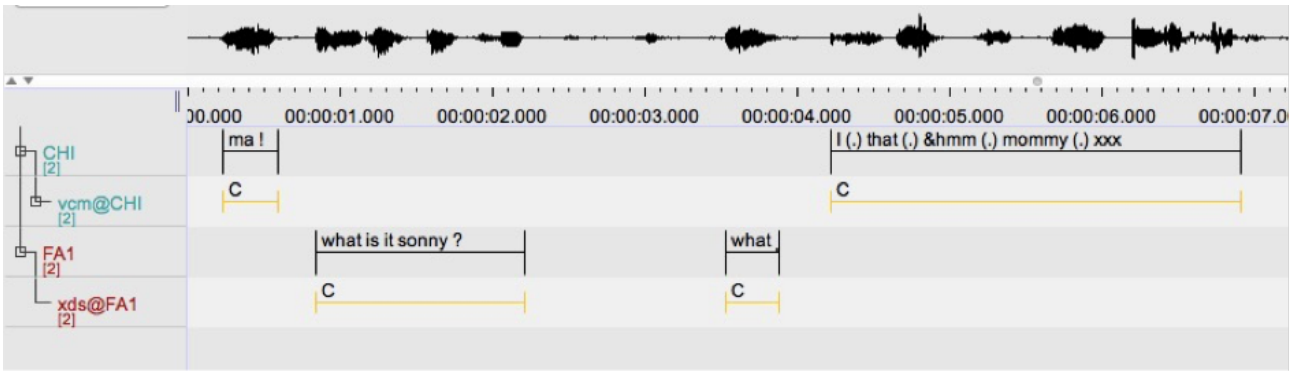

Figure 2: *The same audio segment, now annotated with the ACLEW DAS template (note: the 'lex', 'mwu', and 'nsy' tiers are hidden).*

2. Workflow

2.1. Software

The DAS workflow uses ELAN [2] for manual annotation and for the application of (semi-)automated analysis tools. ELAN is open-source, multi-platform freeware designed for annotating media with language research needs in mind. It is under active development and has infrastructure supporting automated analysis plug-ins. This plug-in feature will be instrumental as we continue building add-ons to perform (semi-)automatic labeling tasks for long-format recordings. ELAN is also interoperable with other commonly used language research software, e.g. Praat [3], and CHAT [1], among others.

ELAN lets users create annotation templates that can be centrally distributed to coders. Templates define standardized annotation tiers and can include closed-set tags and define strict cross-tier relationships (e.g., limiting annotations on one tier given those on another). As such, templates enforce cross-coder consistency. Because each template is modifiable, users can also add tiers onto the core template as needed. At the moment, DARCLE uses Open Science Framework (`https://osf.io/4532e/`) to distribute up-to-date template files, enabling others to freely clone and adapt the basic template to their own needs. We are currently integrating this workflow with a virtual machine framework (e.g., the Speech Recognition Virtual Kitchen [4]) to facilitate tool development as annotation proceeds.

2.2. Data selection

Because DAS is designed to accommodate recordings that are often 10+ hours, there is no assumption that annotations will span the entire recording. DAS instead assumes that researchers will first identify a prioritized subset of their long-format recordings for annotation, however they see fit.

For example, some researchers may want to annotate certain activity types or engagement levels, such as social play, book-reading events, mealtimes, etc. In this case, the first step is to find a process to select those periods of interest. An alternate approach is bottom-up selection, which in its simplest form samples a fixed number of minutes per hour to get frequency estimates for events of interest (e.g., child-directed utterances). If researchers are not interested in specific event-types per se, but only want to analyze ongoing speech, they may combine an automated first-pass to demarcate silence and/or vocal activity with a second-pass that randomly sub-samples from vocally active periods. While DAS works 'as is', we anticipate smooth integration of first-pass vocal activity detection and second-pass sub-sample annotation using ELAN's existing plug-in structure in the near future.

2.3. Deciding what to annotate

The second step is for researchers to determine precisely *what* will be annotated once their sub-samples have been selected. Annotation in ELAN is implemented in "tiers" of codes that are time-aligned to a media file in which the tiers can relate strictly (or loosely) to each other. Importing the concept of "speaker tier" and "dependent tier" from CHAT [1], DAS assumes that each speaker in the audio recording gets a top-level tier, beneath which dependent annotation tiers can be added to supplement top-level information. Top-level tiers can also be used for labeling non-speaker audio (e.g. TV) to better inform the development of new speech processing tools.

2.3.1. Speaker tiers

Speaker tiers serve to demarcate stretches of speech/vocalization by individual speakers. The basic

DAS template is meant for recordings of child language environments, so in its most minimal form it has two speaker tiers corresponding to the two key people we expect to be in most recordings: the target child wearing the recorder (CHI), and a female adult (FA1). As needed, coders can add or edit these "speaker" tiers to include, for example, male adults and other children who are present in the recording. Within the DAS, speaker tiers are labeled systematically with a unique ID that indicates broad-class properties of the speaker: gender (male/female/unknown) and age (adult/child/unknown), plus an integer indicating order of appearance in the present annotation (e.g., FA1, FA2, and FA3 if there are three female adults). This broad-class characterization of each speaker can be expanded through ELAN's metadata capabilities, which allow for associating each speaker tier with additional identifying information. For example, we can add the speaker's relationship to the target child (e.g., mother) or speaker-specific details (e.g., age 34) in a format that interfaces smoothly with CHAT transcription (e.g., "eng|BergelsonCorpus|FA1|34;|female|||Mother|||"). While broad-class speaker information is mandatory for individuated speakers in the DAS workflow, speaker-specific identifying information is optional. Researchers working in either style (broad or specific speaker-tags) will thus have metadata that is CHAT-compatible. As with short-form recordings, a remaining challenge for researchers is deciding when a speaker's role is meaningful enough to deserve a speaker tier (or to include annotations of their speech, if they are backgrounded). DAS remains agnostic about this—it should be driven by the research question at hand.

After considerable piloting, we propose that each annotation on the speaker tier should be an utterance in the case that it is a linguistic vocalization; in the case of non-linguistic vocalizations (e.g. laughter), the annotation interval should span the whole vocalization (including short inter-vocalization intervals). As in CHAT, DAS speaker tiers annotations are built to contain orthographic transcriptions. However, transcription may not always be necessary or feasible; the focus on labeling segments in tiers, rather than transcription is a key difference from existing systems. If forgoing transcription, DAS speaker tiers may be filled with place holders ("0.") to indicate a non-transcribed utterance available for future coding. When coders *do* make orthographic transcriptions, DAS stipulates minCHAT formatting, i.e. the *min*imal formatting requirements needed to ensure compatibility with the CHAT-CLAN system [1].

To review, a basic implementation of the DAS will result in vocalization intervals for the set of speakers determined to be relevant by the researchers (Figure 2).

2.3.2. Dependent tiers

In addition to these basic speaker tier annotations, researchers might often be interested in adding other information about utterances. Borrowing again from the basic concepts used for CHAT, every DAS speaker tier can have any number of *dependent* tiers that contain additional information about the utterances on the parent tier. Dependent tier annotations in DAS are typically symbolically linked—they inherit onset and offset values from higher-level annotations—so that hierarchical annotations can be nested (e.g., utterance–word–morpheme–phone). We illustrate this with one large multi-lab project using DAS with standardized dependent tiers: Analyzing Child Language Experiences around the World (ACLEW; `https://osf.io/kex23/`).

ACLEW is a T-AP Digging into Data-funded project bringing together investigators from 9 labs in 6 countries. It aims to combine raw long-form recordings from several languages and cultures with speech technology tools to develop a better understanding of how language is acquired. We use the DAS to create the manual annotations needed to fuel tool development. The ACLEW DAS template is the first major test of the DAS and will help to streamline the running process between manual annotation and (semi-)automated tool development and application. For present purposes, it also illustrates how researchers can use dependent tiers to tailor the DAS to their needs.

Non-target-child dependent tiers. In the ACLEW DAS template, speakers other than the target child have a single dependent tier that must be manually annotated: 'xds', which indicates the type of addressee (child/adult/both/other) using a restricted set of tags to facilitate cross-lab consistency. Many researchers collecting long-format home audio recordings are interested in characterizing the child's language environment and, in particular, in quantifying the amount of (child-directed or overheard) speech that the target child hears. In the short term, these annotations give us insight into the form, quantity, and distribution of speech actually addressed to the child. In the long term, they will also allow us to develop and hone acoustic and linguistic models that classify novel vocalizations probabilistically into these addressee categories, along the lines of our 2017 ComParE sub-challenge [5].

Using orthographic transcriptions from the speaker tier level, we will also automatically generate an additional dependent tier: 'nsy', which indicates the number of syllables present in linguistic vocalizations. This can be done by either using a look-up dictionary or forced-alignment of the utterance contents to the audio, both viable via plug-in (see, e.g., PhonBank [6]). This interplay between partly manual and partly automated tools within an easy-to-use infrastructure is exactly what the DAS aims to facilitate. In this case, we can use transcription to extract and train syllable recognizers in the hope of further improving upon these methods and other approaches currently in use (such as LENA's use of a speech recognizer to count consonant and vowel sounds), bringing us closer to faithful and comprehensive descriptions of children's linguistic input that are currently out of reach.

Target child dependent tiers. For the target child (CHI), the ACLEW DAS template uses a cascade of manually annotated dependent tiers: 'vcm' (vocal maturity), 'lex' (lexical utterance), and 'mwu' (multi-word utterance).[1] Like the 'xds' tier, each of these dependent tiers comes with a restricted vocabulary to facilitate cross-lab consistency. The ACLEW workflow for child speech expects annotators to first add the vocal maturity of each vocalization, e.g., crying, non-canonical syllables, and canonical syllables. Vocal maturity tags give us insight into children's productive communication abilities when transcription has little to offer or is prohibitively difficult, and are defined in a language-independent way, thus allowing unbiased cross-linguistic and cross-cultural comparisons. In the same style as the syllable recognizers mentioned above, we hope to eventually train classifiers to tag uncoded portions of our datasets [7]. Implementation within a coding-training-application environment (like ELAN + DAS) facilitates this process.

The vocal maturity codes can also be used to speed up annotation on the other two tiers: 'lex' and 'mwu'. Canonically

[1] In practice, some of these tiers are age-dependent; vocal maturity is only coded until fluent lexical speech emerges (around 18 months), the lexical tier is only coded after word-production begins, and multi-word utterances are only coded after children begin to combine words.

babbled vocalizations are tagged for whether they include a recognizable lexical word and, if so, whether they include multiple recognizable words (a binary decision in both cases). In addition to the linguistic analyses we can perform on the manually annotated data, the verified lexical segments allow us to assess the accuracy of off-the-shelf speech recognizers on child language data. Our expectations for these recognizers are low at present, but optimistic given the growing rate of advances in speech technology. Building better training data for these recognizers is fundamental to their future development, and DAS helps ACLEW take a step in that direction. Big data descriptions of child language development are ripe for improvement. For example, LENA's software uses a recognizer trained on English newscaster speech (arguably an ill-fitting training corpus for child speech environments) to find consonants and vowels in children's vocalizations. Given that we can extrapolate language-specific qualitative descriptors of vocal development from these vocal maturity measures, improving their automated accuracy is critical for research progress. An initial step for ACLEW is to generate an additional 'nsy' tier to get information about infants' syllable production, as a proxy for vocal maturity (as with non-target-child speech, described above).

In short, the ACLEW template illustrates a number of the desirable properties of a DAS-style workflow: users can define and elaborate annotation tiers with specific research questions in mind in a way that facilitates the efficiency and comparability of annotators' work and is forward-looking with respect to integration with current and future algorithmic approaches.

2.3.3. Tips for implementation

How does annotation actually proceed given an audio sample (e.g., 5 minutes) and a template? We recommend that coders work cyclically across their data, choosing a manageable unit of time (e.g., 1 minute) in which they add all the annotations needed across their tiers before moving on to the next time unit, i.e., inching across their sample, one time-unit at a time. Within each cycle, we have found it most efficient to annotate on one level at a time, focusing first on parent tiers and then tracing the existing hierarchy of dependent tiers down until all annotations have been filled for a time-unit. Further information and more tips can be found in the DAS training materials, available on OSF (`https://osf.io/4532e/`).

3. Annotation exercises

We asked seven annotators with varying linguistic expertise to self-train on our online materials and implement the ACLEW DAS on a shared 5-minute audio clip. Annotation took approximately 2–3 hours, two-thirds of which was typically used for utterance segmentation. The resulting annotations showed high agreement on utterance boundaries ($\sim$80% utterances with full consensus; $>$90% with majority consensus), and excellent performance on some dependent tier annotations (99% of adult utterances had a 0.75+ majority annotation across coders), but not others (34% accuracy on vocal maturity, using one expert's judgments as a gold standard). Such exercises are crucial for refining the specific implementations of the DAS workflow used to run large projects and the ACLEW implementation is still actively under development (see OSF site for more details).

4. Conclusions

We introduce DAS, a barebones annotation workflow designed with long-format recordings and (semi-)automated annotation tools in mind. A key feature of DAS is flexibility in implementation; there is enormous cross-researcher variation regarding how to sample data for annotation, which speakers or speech/interaction types are relevant, and in what detail to complete their annotations and metadata. Rather than introducing rigid requirements for annotation content, we take an approach of providing an infrastructure and communal documentation sources to encourage annotation compatibility across corpora collected, annotated, and analyzed with specific researchers' questions in mind. The only rigid aspects of this scheme are in the naming and relationship types between tiers and in the use of speaker tiers as the top-level tier (which is necessary to maintain interoperability with CHAT and CLAN). These aspects are in-line with basic annotation requirements used nearly universally across child language corpora.

The DAS seeks to provide researchers with an easily customizable workflow that will ultimately result in annotations that are useful for the community at large. To this end, it complements data repositories (e.g., HomeBank) researchers are already using and the formats associated with them (e.g., CHAT) such that DAS templates, coding documentation, and tools are useful, accessible, and harmonize with researchers' own annotation needs. Collaborative DAS-style annotation across diverse child language corpora will be indispensable for improving the accuracy and ease of application of (semi-)automatic tools—tools badly needed to generalize our understanding of children's linguistic input to their full daily experience. In the meanwhile, we anticipate that the accumulation of common annotations across long-format corpora will inspire the same kind of data inspection and re-use of hard-to-collect data that CHILDES has enabled over the past three decades; only now with long-format recordings. We hope that this too will attract researchers' interest in engaging with the DAS. Community building and working with large networks of researchers like DARCLE is at the heart of this process; pooling resources at all levels will increase our ability to put to meaningful use these new and challenging, but incredibly rich and ecologically-valid long-format recordings.

5. Acknowledgements

This work was funded in part by NWO Veni Innovational Research Scheme 275-89-033 (MC), NIH DP5-OD019812 (EB), SSHRC Insight Grant 435-2015-0628 (MS), and the Agence National de la Recherche (ANR-14-CE30-0003 MechELex, ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC; AC). The authors further thank the DARCLE group for their participation in discussion and exercises surrounding the development of DAS, and Brian MacWhinney for advice on CHAT compatibility.

6. References

[1] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[2] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: A professional framework for multimodality research," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 2006.

[3] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer

[computer program]," 2013, version 5.3.51, retrieved 2 June 2013 from http://www.praat.org/.

[4] F. Metze, E. Fosler-Lussier, and R. Bates, "The speech recognition virtual kitchen," in *Proceedings of INTERSPEECH 2013*, 2013.

[5] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Proceedings of the Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, in press.

[6] Y. Rose, T. Schmidt, and K. Wörner, "Multilingual phonological corpus analysis: The tools behind the phonbank project," *Multilingual corpora and multilingual corpus analysis*, pp. 365–381, 2012.

[7] E. H. Buder, A. S. Warlaumont, D. K. Oller, B. Peter, and A. MacLeod, "An acoustic phonetic catalog of prespeech vocalizations from a developmental perspective," *Comprehensive perspectives on child speech development and disorders: Pathways from linguistic theory to clinical practice*, pp. 103–134, 2013.