Vetting Manual

Preparation of Recordings for Unrestricted Publication in HomeBank

Version 1.1

Mark VanDam,[1] Anne S. Warlaumont,[2] Brian MacWhinney,[3] Melanie Soderstrom,[4] & Elika Bergelson[5]

[1]Washington State University

[2]University of California, Merced

[3]Carnegie Melon University

[4]University of Manitoba

[5]Duke University

June 18, 2018

Available electronically:
https://homebank.talkbank.org/pubs/vetting.pdf
https://doi.org/10.21415/T56H4M

1 Purpose

The goal of vetting daylong audio files is to produce a redacted audio file that can be disseminated in an unrestricted, open-access fashion with sufficient expectation that the audio is free from content not intended to be public. A challenging task is to determine precisely how "not intended to be public" is interpreted and put into action. This document is part of the HomeBank project (http://homebank.talkbank.org/). Note that for special populations, such as at-risk populations, or indigenous populations with little experience with the internet, special considerations may apply. The following is not intended as a set of general purpose guidelines for data-sharing, but rather as a process for those files which have already been deemed low-risk for sharing and for which appropriate consent has been obtained from the participants.

Vetting requires that the entire audio record is listened to by a trained judge whose job is to identify and mark segments of the audio record that should not be made public. At this time, we do not recommend automated vetting or vetting based on the results of automated transcription, though of course automation techniques can be usefully employed to enhance the transcription or reduce the time to transcription. There are two general approaches to vet an audio file. The first approach is vetting-only, in which a judge identifies only those segments to be vetted but does not annotate or transcribe the audio record in other ways. The second approach is veting-while-transcribing, either simultaneous with original transcription or on an extant transcription. The second approach is of course more time consuming, but may be used in conjunction or simultaneously with other project goals.

Assuming the user has obtained an audio recording suitable for vetting, this document describes the procedure of file preparation and the procedure for vetting unwanted information from that file. When the vetting is complete, the audio is processed to formally redact the marked segments, and the audio can then be made publically available.

2 General Procedure

This section describes file preparation for the vetting procedures.

2.1 *File preparation from recordings alone*

If you are transcribing and/or vetting anew—that is, starting with (only) an audio file—you may use the general procedures described on the CLAN (http://childes.talkbank.org/clan/) and the CHILDES (http://childes.psy.cmu.edu/clan/) websites. If you using an alternative annotation approach, such as the DARCLE annotation scheme, please ensure that your basic formatting allows for conversion to the format described below (in CHAT) so that the audio may be scrubbed using your annotated files. This will facilitate more straightforward inclusion into the HomeBank archive.

2.2 *File preparation using LENA recordings*

After a LENA recording is obtained and processed, use the LENA software to generate the WAV and associated ITS diarization file. To generate the recording, complete the following steps.

a. Select *LENA Reports*
b. Select *Export Data*
c. Select desired participant and dates
d. Select *Recording to TRS/CHA* (disregard the TRS or CHA file created[1])

To generate the ITS file, complete the following steps using LENA software.

a. Select *LENA Reports*
b. Select *Export Data*
c. Select desired participant and dates
d. Select *ITS*

Using an up-to-date version of CLAN, process the ITS file using the CLAN software and the command *lena2chat*. The *lena2chat* command will use the file and directory structure in the preparation of the files, so it is recommended that you put the WAV and ITS files into a directory structure with four-character names as the directory node immediately containing the file to be processed. This step produces results that are consistent with the conventions of the HomeBank database.

To run *lena2chat*, locate your working directory above the files to be processed and type "*lena2chat* +re *.its" (without the quotes) into the command line. The flag "+re" will recurse the command, processing all the ITS files in that directory tree. The processed CHA file will be in the same directory as the ITS and WAV files. The *lena2chat* command names the output CHA file a six-digit number "ID_YYMMDD.cha" indicating the child's ID (assumed to be the same as the name of the folder containing the file) and age of the child at the recording. The *lena2chat* command will also rename the WAV file contained in the same folder to the same name as the CHA file with the *.wav* extension.

2.3 *Using CLAN with CHAT files imported from LENA*

Using an up-to-date version of CLAN, open the CHA file and associated media (see section 2.1 above if necessary). The WAV file should be in the same folder as the CHA file. To

[1] The lena-generated CHA file excludes information that the *lena2chat* conversion in CLAN preserves.

begin listening to the audio, press *Esc-8* or use the drop down menus. As the audio plays, a black bar will highlight the current line being played. The user can stop the audio at any time by clicking in the active CLAN window. When the user hears content to be vetted, the audio can be stopped, then the user returns to the line that contains the information to be vetted. For vetted utterances, enter the postcodes [+ cut1] or [+ cut2] (see also section 3.2 below) on the line to be vetted; the postcode must be followed by a space, and it is placed at the end of the utterance after the final delimiter but before the bullet marking the time values, as in the example below.

```
*CHI: my uncle is John Doe from Chicago. [+ cut1]
```

Do not delete anything that is already existing in the CLAN file. Audio playback can be resumed at the edited point with F6 or the drop-down menus.

2.4 *Notes on using CLAN*

Make sure you are "checking" the file periodically. To do this in CLAN go to *Mode* > *Check opened file*. CLAN will report technical or syntax errors if present. If errors are identified, correct and continue.

Mode > *show line numbers* will number each line. Turn this on to facilitate resumption of a task or to make notes about specific line numbers.

Edit > *Go To ...* will take the user to a specific line number.

Be sure to use F6 to listen to the files and not F5, which inserts segments.

3 Vetting procedure

The purpose of vetting is to tag or annotate utterances that are deemed unsuitable for publication or public dissemination. The goal of vetting is not only to eliminate obviously personally identifying details (such as full names or addresses) but also to eliminate any content that may or has the potential to embarrass, shame, dishonor, discredit, or defame any person, including the talker or the subject of a segment. Spoken episodes or exchanges have linguistic and social contexts, which should be taken into account. For example, the phrase “I could just kill you right now” could be uttered in anger or in jest. Contextual disambiguation is critical in deciding whether a segment should cut or retained.

It is up the user whether or not to transcribe those segments that are tagged to be cut. That is, if a segment is determined to be cut, it may not make sense to memorialize the literal content of the objectionable segment. In future editions of the vetting manual, or at the discretion of the transcriber, a vetted segment may be transcribed with a placeholder for content such as “[argument between adults]” as opposed to the (literal) content of the argument itself.

The following sections describe the vetting procedure.

3.1 *Usage*

Use postcodes to mark utterances to be cut from the public file. Postcodes are used in square brackets with a plus sign, "+", placed at the end of the utterance. Postcodes are symbols placed into square brackets at the end of the utterance. They should include the plus sign and a space after the left bracket. There is no predefined set of postcodes. Postcodes apply to the whole utterance (as opposed to scoped codes). Notice the CHAT syntax in the second and third examples below making use of the *explanation* and *comment-on-main-line* notation.

Example: `*CHI: his birthday is December 7, 1941 ? [+ cut1]`

Example: `*FAT: [= vetted remarks] ! [+ cut2]`

Example: `*FAT: [% vetted content] .  [+ cut1]`

3.2 *Labels in CHA file*

There are two codes labels used for vetting, *cut1* and *cut2*. The *cut1* label is used when the decision to exclude that segment is unambiguous and requires no additional human input. The *cut2* label is used when the decision to cut is likely but unclear to the individual transcriber. In the case of a *cut2* label, typically several transcribers or a research team re-evaluates and forms a consensus. The *cut2* label is likely to be used rather liberally for most transcribers. Description and examples are given below.

cut1 – Cut without reservation; explicit personal identification (full name, phone number, address, SSN, DOB, etc.), a private episode, or personally embarrassing statements.

Example: "she lives at 123 Main Street in Tulsa"

Example: "right now I'm at the corner of 8th and Oak near my house in Dallas"

An example of an intersection may be indentifiable in certain contexts such as a small town, but may be sufficiently unlikely to be identifiable, such as a densely populated area of a large city, that the decision not to cut may be warranted.

Example: "Samantha Rae Jones, go to your room"

Example: "Kaitlyn Marie Smith, come see Grandma"

Example: "his birthday is October 1, 2005"

If this is the birthday of the child and that information is not otherwise redacted, the information may be entailed elsewhere in the database and cutting here may be redundant. On the other hand, this might be in reference to an adult, an unidentified person, or without sufficient context that a specific person would likely be identified.

Example: “let’s go smoke this joint where no one can see us”

Example: “I can not stand Susan’s shrill voice”

Example: gossip about a co-worker or neighbor

Example: “Hi. Yes. I’d like to schedule a doctor's appointment. I can no longer feel my IUD. Can I get an appointment at the Simpson Road clinic?”

Since this refers to reproductive organs and reveals medical information, this likely would be *cut1*. Additionally, this may be a uniquely identifiable location, “Simpson Road.” On the other hand, this may be an unidentified interlocutor on the recording that offers only that segment (such as in an overheard conversation) without context, possibly justifying retention of that segment. In the latter case, labeling this line as *cut2* would flag it for discussion.

cut2 – Borderline case, should be discussed and reviewed with multiple team members.

Example: name and location of employment. “Jason, my husband, works at Luigi’s”

The decision to cut might be made if “Luigi’s” is deemed unique or may be used in context to identify an individual, perhaps as the only establishment of that name in a small town identified in another section of the recording. On the other hand, the research team may deem “Luigi’s” to be sufficiently common that it need not be cut, perhaps due to several similarly named establishments in the town otherwise named in context.

3.3 *What not to cut*

3.3.1 ***Names of institutions.*** Names of institutions are retained, provided it is not a unique name of institution that would foreseeably identify a specific person. For example, the name of a university would not be vetted, but a small, private preschool in a specific city may be.

Example: identifying the name of a professor and university alone would not likely be vetted. “I’m going to Voice Disorders taught by Professor Jackson”

In this case, the context should be taken into account. If the dialogue identifies a specific person and institution (as might be possible with Professor Jackson above), that information is not private and would not likely be marked for deletion. If, on the other hand, the dialogue included information that would be expected to be private and identified a specific person (for example, and ad hominem comment about Professor Jackson who works for a specific university), that segment would likely be marked for deletion.

3.3.2 ***Personal Names.*** Personal names are retained.

Example: “tell John to come over to my house”

Example: “I’m going to bring the kids to Aunt Calliope’s ranch”

The combination of specific details such as a city (e.g., *Denver*) with a less common name (e.g., *Enid*) may be deemed to increase the potential to identify an individual, and those segments may be considered for deletion.

Example: “because Enid and I live in Denver, we are Broncos fans”

3.3.3 ***City names.*** City names are retained, provided they do not uniquely identify an individual.

Example: “yes, we live in Spokane, Washington”

3.3.4 ***Scatological and body functions.*** Scatological and bodily functions are retained, provided that a reasonable amount of discretion is used to limit potential embarrassment of any individual in the recording, explicitly identified or otherwise. Human biological bathroom (or bedroom) functions without associated identity and a certain amount of ambiguity are likely not too revealing. Similarly, tub-filling or general bathroom use, including conversation, is appropriate to keep.

Example: “I go poopy in the potty”

Example: “Daddy wipe my bottom”

These guidelines may be interpreted differently in other cultures or environments, and exceptions to these guidelines, either more liberally or more cautiously applied, are at the discretion of the researcher and research group.

3.3.5 ***Obscenity and profanity.*** Obscenity and profanity are retained, provided those segments would not unduly embarrass a reasonable person.

Example: “she’s fucking funny. Listen to her say her own name” (unvetted)

> Example: [from a parent directed to a child] “that's why you need to start putting your damn hair up instead of fighting me all the damn time [be]cause you think you're the fuckin’ boss.”
>
> > Here, this is probably cut without more context. If it is accompanied by laughter or lightness in context, maybe it is not vetted; if it is said in anger or in obvious contempt, for example, it should be vetted.
>
> Example: “if I fucking hear *one more* fucking word I’m going to fucking *kill* you!”
>
> > Here, vetted on account of anger of the talker, but not on account of obscenity.

3.3.6 ***Arguments.*** Arguments between adults are retained if they are considered minor in nature. If an argument is heated and likely to embarrass or otherwise implicate any person (identified or otherwise), it should be vetted. When an adult loses his/her temper, and this anger or aggression is directed toward a person (even a person not present).

> Example: “have you looked at this floor? Why are you watching a pregame; it’s not even a real game! Stop being so lazy and help out for once!” (unvetted)
>
> Example: “you’re late again! I don’t appreciate that!” (unvetted)

3.4 *Considerations*

In the case of uncertainty with respect to whether to mark as vetted/cut or at what level, a note of it should be made and other members of the research team should be consulted.

3.5 *Additional procedures*

Some research teams may choose to make a note on a separate document of the line number and content of vetted/cut material to facilitate easy assessment. This supplemental document can be a word-processor document or entries in lines of a spreadsheet. Record the line-number of the tag or annotation, the transcribed content of the line, the “cut” code, and any notes/comments.

3.6 *Processing the finished product*

Once the vetting is complete, the full, unredacted audio, accompanying CHA file, and external record of vetted segments (if present) are processed to formally redact the audio segments marked to be vetted. After the automated process deletes the marked segments, the modified audio can be hosted in the HomeBank database, including the portion of the database available for unrestricted public access. The transcriptions of the vetted materials in both the CHA record and (if present) an external file will be deleted and will not be available in the database.