HLVC Anonymizing Project

Protocol created by Julia Petrosov, 17 April 2020
Edited by NN 20April 2020 – added minor details for continuity.
Edited by JP 05 May 2020 to account for .eaf files and minor details.
Edited by NN 09May2020 to account for untranscribed files.
Edited by NN 15May2020 from HLVC meeting suggestions (Save .eaf as… , aligning tokenized words).
Edited by NN 20May2020 from workstudy questions (files that don't need edits, filename details)

For this project, we have the following priority order (for Heritage and Homeland samples):

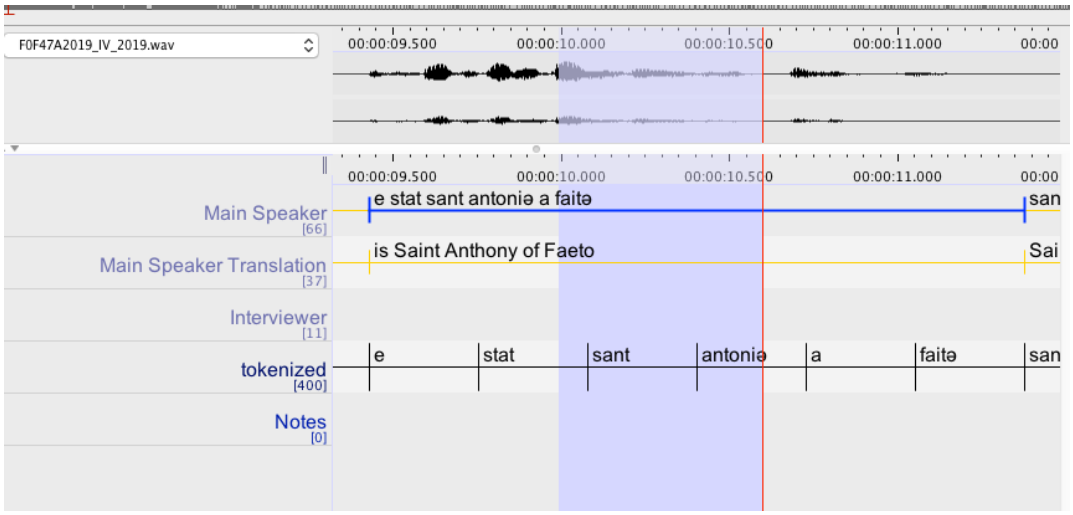
1. Already-proofread IV and FW .eaf+.wav
2. Transcribed IV and FW .eaf+.wav
3. EOQ .wav files
4. Untranscribed (or incomplete) IV and FW files that we need to complete our sample. (Note that, for some languages, we have many extra files that we don't expect to ever use. These should be marked as such in the catalog. We don't need to anonymize them.

Always check to make sure that you use the newest version of a particular .eaf file, by using the Finder Search function within your whole LANGUAGE folder. (If working remotely, use "Remote search" in FileZilla. It's the binoculars icon at the top of the window.)

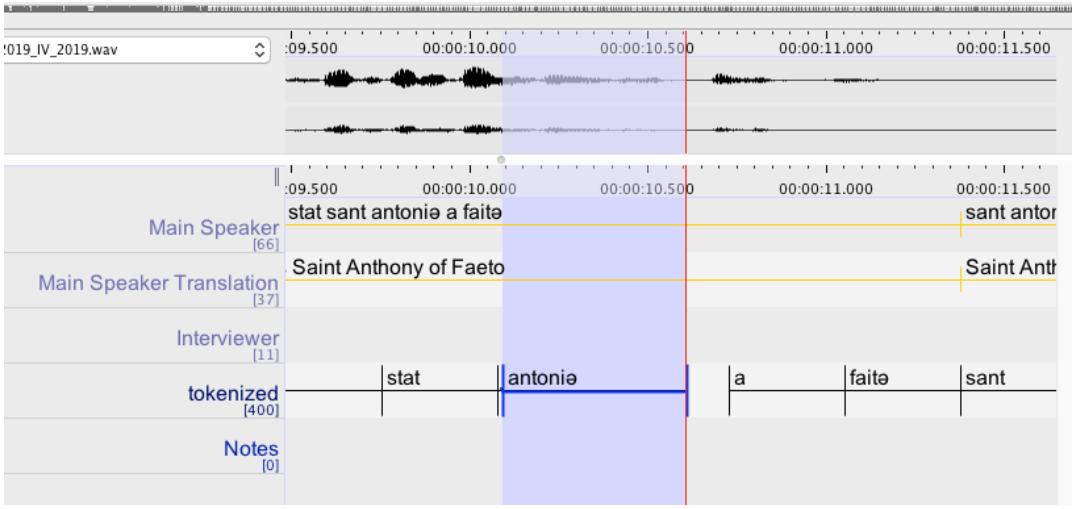
For files where an .eaf already exists:

1. Save a new copy of the .eaf with "_anon" added to the end of the filename.
2. Open the new .eaf.

3. In the Tiers menu, choose "Tokenize Tier..." Click the "Create New Tier…" button in the window that pops up. Name that new tier "tokenized." Click "Start." A new tier will appear with each word in a separate annotation field. Note that words are exactly not time-aligned to the right part of the sentence. In this example, I've highlighted the part of the soundwave where the name "Antonio" is actually said. It doesn't line up with the transcription of that word in either the "Main Speaker" sentence transcription nor the "tokenized" word transcription tier.



4. Adjust the boundaries (Option + click and drag on boundary) of that single-word annotation until it matches the audio. Listen to make sure. (The blue triangle button with the "S" label plays exactly what you've selected.) Here is the adjusted version, again with the word "Antonio" highlighted in the soundwave.

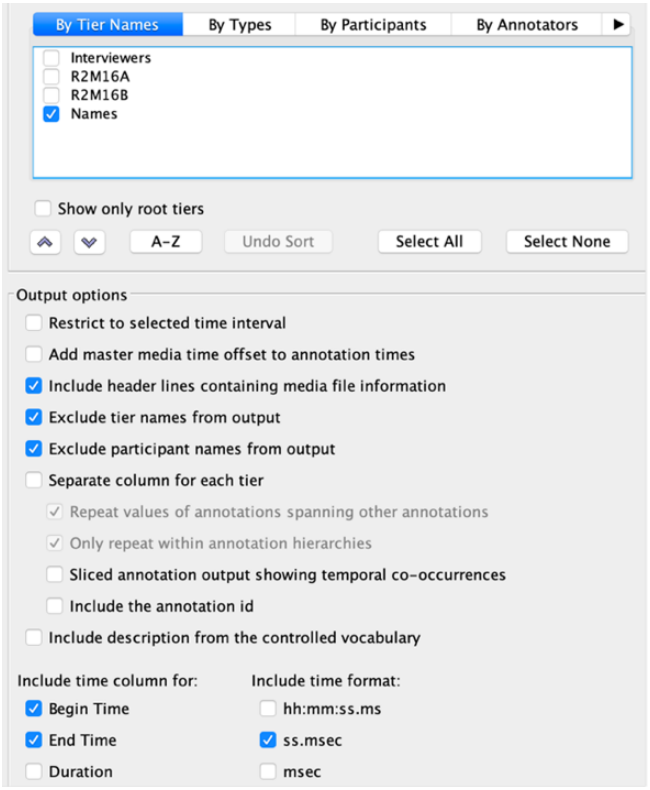


5. Create "Names" tier with controlled vocabulary: "Omit" and "Unsure." This gives you a pull-down menu to mark each identifier's timeslot with a label "Omit." (Use "Unsure" when you want to check later with someone if a word should be omitted.)
    a. Edit>Edit Controlled Vocabularies (or copy with FileZilla, then import 128.100.214.80/Users/hlvc/RUSSIAN%20%28Heritage%29/Anonymizing%20Project/IdentifierCV.ecv)
        i. CV Name: Identifier
        ii. "Add" (on the righthand side)
        iii. Entry Values (enter each separately): *Omit*, *Unsure*
        iv. "Add" (below the value box)
    b. Type>Add New Tier Type

i. Type Name: Identifier
ii. Stereotype: Symbolic Association
iii. Use controlled vocabulary: Identifier

c. Tier>Add New Tier
i. Tier Name: Identifier
ii. Parent Tier: Token
iii. Tier Type: Identifier

6. Add annotations in the Identifier tier in any place where you see: Personal last names or super-unusual first names, street names, church names, school names, specific workplaces, and anything else that may be used as an identifying factor. Check with Naomi about anything you are unsure of. **Make sure the annotations you add line up with the audio meant to be silenced -- the text of the tokens is not always in line with the audio.**
    a. Last names should be represented by the first initial only, e.g. "Naomi N.".
    b. Birthdates, if mentioned, should have the day omitted, e.g., "April XX, 1967"
7. Replace all the written identifiers with *[REDACTED]* in the tokenized and Main Speaker tiers.
8. Use FileZilla to put the edited file, "_anon" added to the end, in the "Anonymized Files" folder corresponding to your language on the appropriate iMac. If your language does not have that folder, create it.
    a. *Note*: If you find that there were no identifiers in the whole file, create a renamed copy of both .eaf and .wav, as in the instructions below, and save to the "Anonymized Files" folder anyway, ensuring we end up with a complete set of files.
9. Export the ELAN file as a tab-delimited text (File>Export as> Tab-delimited text). You should export it using these options:



10. Save the .txt file with the Speaker code in the filename as an identifier, e.g., SPEAKERCODE_anon.txt.
11. Open the same .wav file with Audacity.
12. From the File menu, select Import>Labels> select your SPEAKERCODE_anon.txt file
13. On the tier with the names, click "Select."
14. Edit>Labeled audio> Silence audio. All of the annotated parts should now be silenced.
15. Save the .wav file and specify anonymity completion in filename. i.e.; R1M45A_IV.wav → R1M45A_IV_anon.wav
16. Place the NEW file in the folder "Anonymized Sound Files" on the iMac, in the LANGUAGE folder corresponding to your language (e.g., iMac1/RUSSIAN/ Anonymized Sound Files. (Use FileZilla.) If your language does not have this folder, create it. The folder should be inside the "Anonymized Files" folder.
17. Update the catalog for your language to specify that the file you were working on has been anonymized. If needed, add a column in the catalog for your language for this, and include your initials and completion date.

For files that have not yet been transcribed:

1. Create a properly-named .eaf file to go with the .wav, if necessary.
2. Create a Speaker tier. Roughly/Quickly transcribe only the clauses containing an identifier, using “[REDACTED]” instead of any identifiers.
3. Then begin with Step 1 above to create the Identifier tier and follow the regular process.

Notes:

- If there is an identifying factor uttered by the interviewer, create a tier for the interviewer using the same CV and tier type, and remember to select both identifier tiers when exporting as tab-delimited text.