

Training Module

How to Build a Reference Database



This tutorial shows you how to build your own SALT reference database using the Research version of SALT. Once built, your database can be used within SALT the same way you use any of the built-in databases.

Steps to Build a Database

1. Prepare the transcripts
2. Organize the transcripts into databases and groups
3. Build the database
4. Test the database
5. Edit the default database setting

There are several steps to building a database from your own transcripts.

First, prepare the transcripts and organize them into databases and groups. Then build and test the database. Finally, change the default database setting. Each of these steps will be covered in this tutorial.

Step 1: Prepare the transcripts

- Required
 - no transcript entry errors
 - same target speaker (1st or 2nd)
 - plus line(s) with participant age and/or grade in school

+ CA: <age in years and months, or age in years>

+ DOE: <date of elicitation>

+ DOB: <date of birth>

+ Grade: <P, K, 1, 2, 3, ...>

The first step is to prepare the transcripts to be added to the database.

Make sure the transcripts do not contain any transcript-entry errors. Transcripts with errors cannot be added to the database.

All transcripts should have the same target speaker: either 1st speaker or 2nd speaker, determined by the \$ speaker line at the beginning of each transcript.

Step 1: Prepare the transcripts

- Required
 - no transcript entry errors
 - same target speaker (1st or 2nd)
 - plus line(s) with participant age and/or grade in school

+ CA: <age in years and months, or age in years>

+ DOE: <date of elicitation>

+ DOB: <date of birth>

+ Grade: <P, K, 1, 2, 3, ...>

All transcripts must contain the participants age, grade in school, or both. This information is included as plus lines at the beginning of each transcript. The participants age can either be entered on the +CA: entry or calculated from the +DOE: and +DOB: entries. The grade in school must be entered on the +Grade: entry. Transcripts missing both age and grade cannot be added to the database.

Optional Plus Lines

+ Gender: <M, F, N>

+ Context:

+ Subgroup:

+ Ethnicity:

+ Location:

+ Select:

Other plus lines, if included, may be used when adding transcripts to the database. If the +Gender: entry is found in the transcript, the participants' gender is stored in the database and may be used as selection criterion along with age and grade.

Although the Context plus line is optional, the database context is always specified when building a database. You will be warned if the context value in the transcript does not match the database context.

Optional Plus Lines

+ Gender: <M, F, N>

+ Context:

+ Subgroup:

+ Ethnicity:

+ Location:

+ Select:

The Subgroup, Ethnicity, Location, and Select plus lines are also optional and are used to specify groups of transcripts. The next step discusses how these plus lines are used.

Step 2: Organize the transcripts

- All transcripts within the same database should be elicited using the same context and target language
- Transcripts may be organized into groups based on subgroup, ethnicity, location, and +Select (user-defined)

Step 2: Organize your transcripts into databases and groups.

Should you create more than one database from your transcripts? If the samples were elicited using different contexts and/or different target languages, they should be stored in separate databases.

Should you create more than one group within the same database? Transcripts can be added to the database in groups based on subgroup, ethnicity, location, and values found on the +Select plus lines. You would add the transcripts in different groups if you want the option of selecting them based on these groups. Examples follow.

Examples of Organization

- Conversation
 - 1 database, no separate groups
- Narrative Story Retell
 - 1 database, 4 groups based on story
- Persuasion
 - 1 database, 2 groups based on location
- Bilingual (Spanish/English) Story Retell
 - 2 databases based on language, 3 groups in each database based on story
- TNL₂
 - 1 database, 4 groups based on task
- NZAU Conversation
 - 1 database, 3 groups based on location and ethnicity

Several of the databases built into SALT are described here to illustrate how transcripts are grouped by subgroup, ethnicity, or location. All the databases contain transcripts from participants of different chronological ages and/or grades in school. Age and grade are not considered when grouping transcripts because users always have the option of selecting age range and/or grade in school.

The **Conversation** database contains transcripts from different locations. Analysis of the transcripts, however, did not show significant differences based on location so all the transcripts were added as one group.

Examples of Organization

- Conversation
 - 1 database, no separate groups
- Narrative Story Retell
 - 1 database, 4 groups based on story
- Persuasion
 - 1 database, 2 groups based on location
- Bilingual (Spanish/English) Story Retell
 - 2 databases based on language, 3 groups in each database based on story
- TNL₂
 - 1 database, 4 groups based on task
- NZAU Conversation
 - 1 database, 3 groups based on location and ethnicity

The **Narrative Story Retell** database contains transcripts from different locations and elicited using different stories. In this database, the subgroup is the story and there are 4 subgroups, one for each story. Analysis did not show significant differences based on location but did show differences based on story. So the transcripts were only grouped by story. When comparing transcripts to this database, users must select the specific story.

The **Persuasion** database contains transcripts collected in the United States and Australia. The transcripts were grouped by location. When comparing transcripts to this database, users have the option of selecting transcripts from one of the two locations.

Examples of Organization

- Conversation
 - 1 database, no separate groups
- Narrative Story Retell
 - 1 database, 4 groups based on story
- Persuasion
 - 1 database, 2 groups based on location
- Bilingual (Spanish/English) Story Retell
 - 2 databases based on language, 3 groups in each database based on story
- TNL₂
 - 1 database, 4 groups based on task
- NZAU Conversation
 - 1 database, 3 groups based on location and ethnicity

The **Bilingual (Spanish/English) Story Retell** transcripts were separated into two databases – one for Spanish transcripts and the other for English transcripts. Both databases contain transcripts from different locations and elicited using different stories. In these databases, the subgroup is the story and there are 3 subgroups, 1 for each story. Analysis did not show significant differences based on location but did show differences based on story. So the transcripts were only grouped by story. When comparing transcripts from bilingual speakers, the user must first choose the correct database based on the target language (Spanish or English) and then select the specific story.

Examples of Organization

- Conversation
 - 1 database, no separate groups
- Narrative Story Retell
 - 1 database, 4 groups based on story
- Persuasion
 - 1 database, 2 groups based on location
- Bilingual (Spanish/English) Story Retell
 - 2 databases based on language, 3 groups in each database based on story
- TNL₂
 - 1 database, 4 groups based on task
- NZAU Conversation
 - 1 database, 3 groups based on location and ethnicity

The **Test of Narrative Language (TNL₂)** database contains transcripts elicited using 3 narrative tasks. The subgroup is the task and there are 4 subgroups: 1 for each task and a 4th subgroup for transcripts containing all 3 tasks combined. Transcripts were added to the database based on subgroup. When comparing transcripts to this database, users have the option of selecting transcripts containing all 3 tasks, or selecting transcripts containing only one of the tasks.

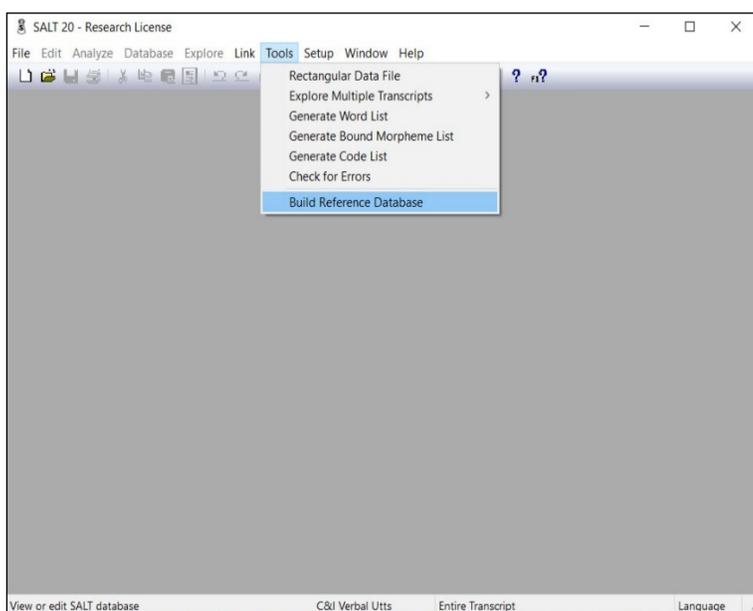
Examples of Organization

- Conversation
 - 1 database, no separate groups
- Narrative Story Retell
 - 1 database, 4 groups based on story
- Persuasion
 - 1 database, 2 groups based on location
- Bilingual (Spanish/English) Story Retell
 - 2 databases based on language, 3 groups in each database based on story
- TNL₂
 - 1 database, 4 groups based on task
- NZAU Conversation
 - 1 database, 3 groups based on location and ethnicity

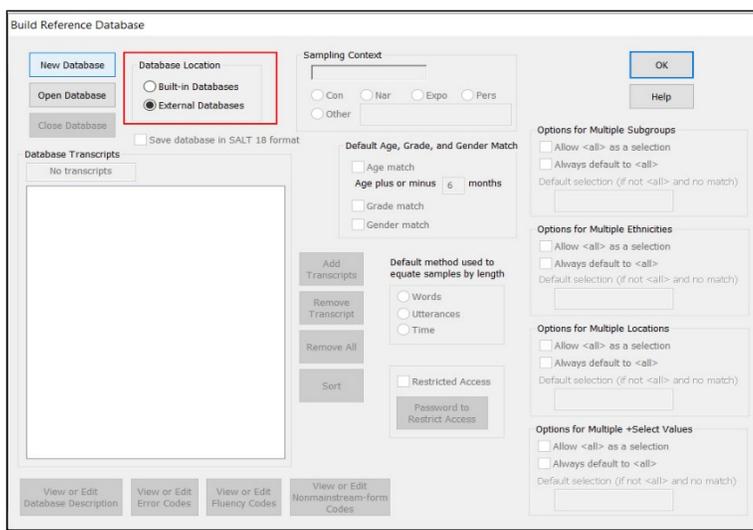
Finally, the **New Zealand – Australia Conversation** database contains transcripts collected in two locations, New Zealand and Australia. A subset of the New Zealand transcripts was collected from participants who were identified as Maori. To provide maximum flexibility, the transcripts were grouped by country and, within New Zealand, further grouped by ethnicity. The transcripts were added to the database in 3 groups: Australia, New Zealand Maori, and New Zealand non-Maori. When comparing transcripts to this database, users must select the location (both locations or specific country) and ethnicity (all ethnicities or just Maori).

Step 3: Build the database

Ok. The prep work is done. It's time to build the database.

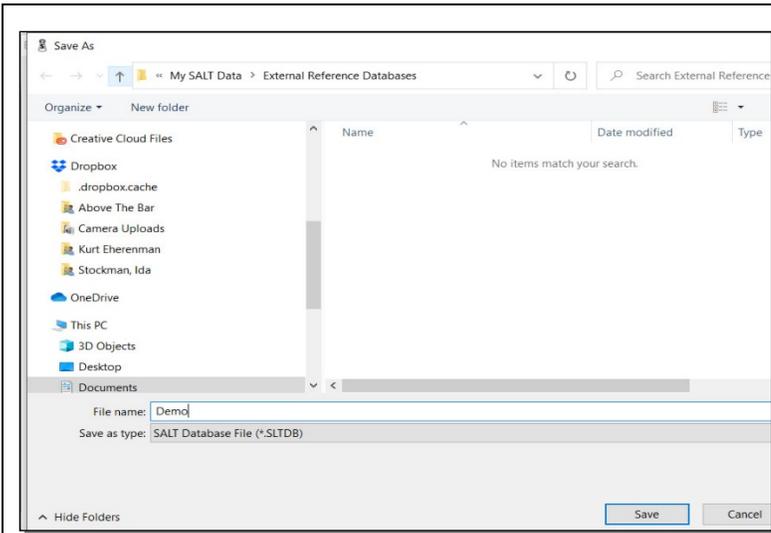


Select "Build Reference Database" from the research Tools menu.

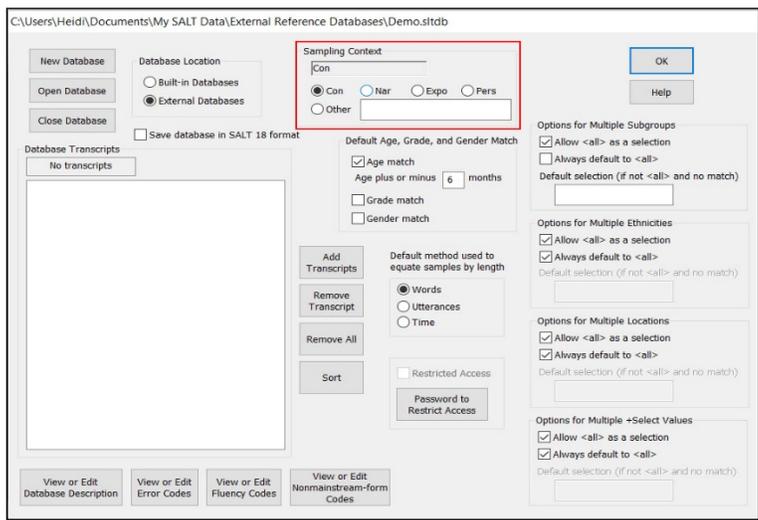


There are 2 places SALT looks for databases. By default, the built-in databases are stored in the ProgramData folder (Windows®) or in the same folder as the SALT program (Mac). And the external databases, such as this one, are stored in the "My SALT Data" folder within the "Documents" folder. These default locations can be changed in the Setup menu.

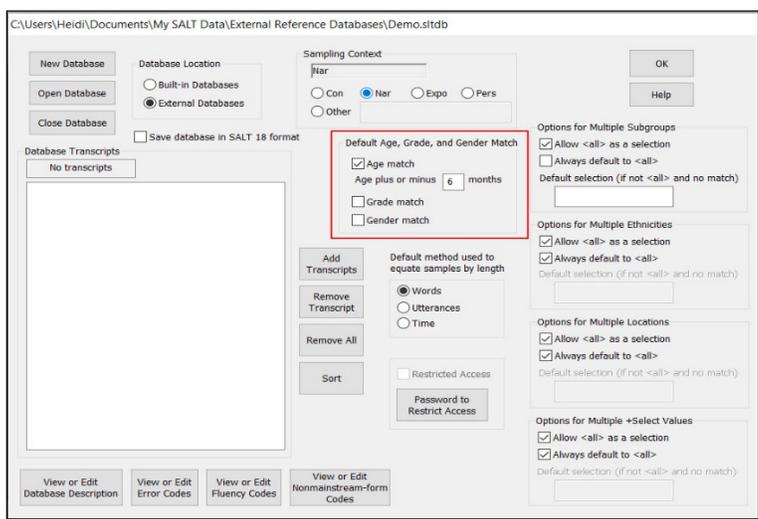
The "Database Location" is "External Database". I'm going to select "New Database".



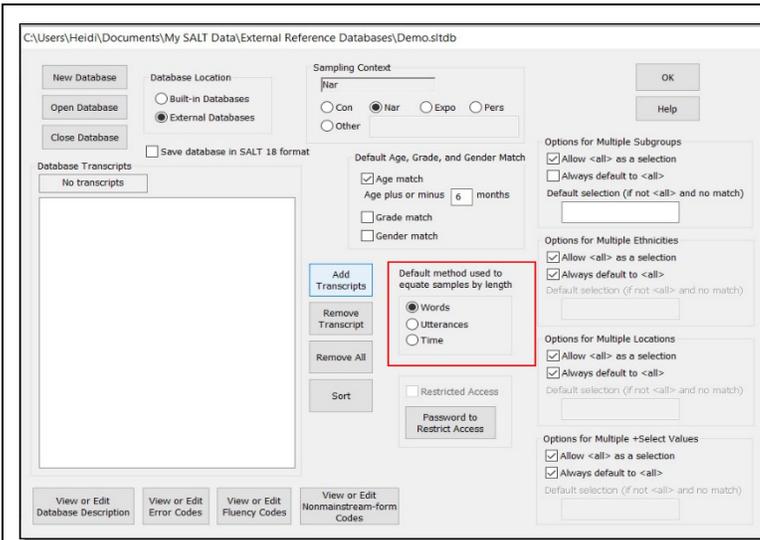
I'm naming this database "Demo" and saving it in the default folder for external databases.



The first setting is the sampling context. For this demo, I'll be creating a narrative story retell database with 2 groups based on 2 different books. So I'm going to change the context to "Nar".



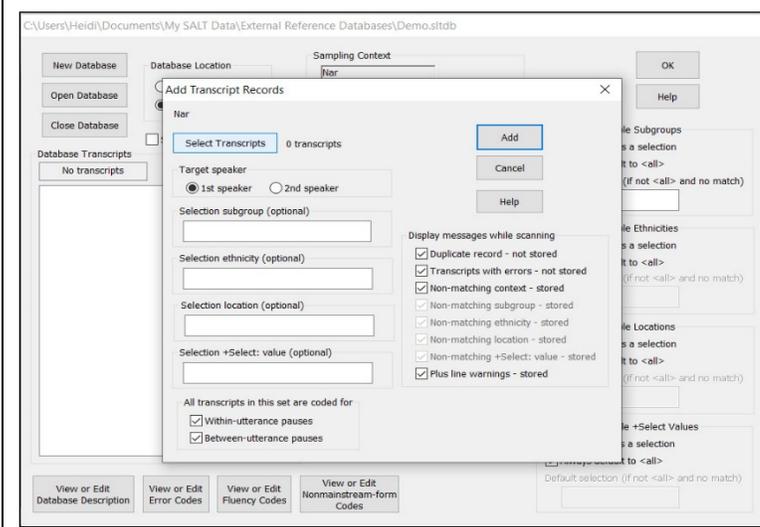
The next section specifies the default settings when users compare their transcripts to transcripts selected from this database. Inexperienced users rarely change default settings so this is an important decision. Most of the SALT databases use "Age match plus or minus 6 months", trading a wider age range with a larger number of matching samples. The bilingual databases built into SALT, however, contain so many transcripts that the default is "Age match plus or minus **2** months" **and** "Grade match". For this demo, I'm keeping "Age match plus or minus 6 months".



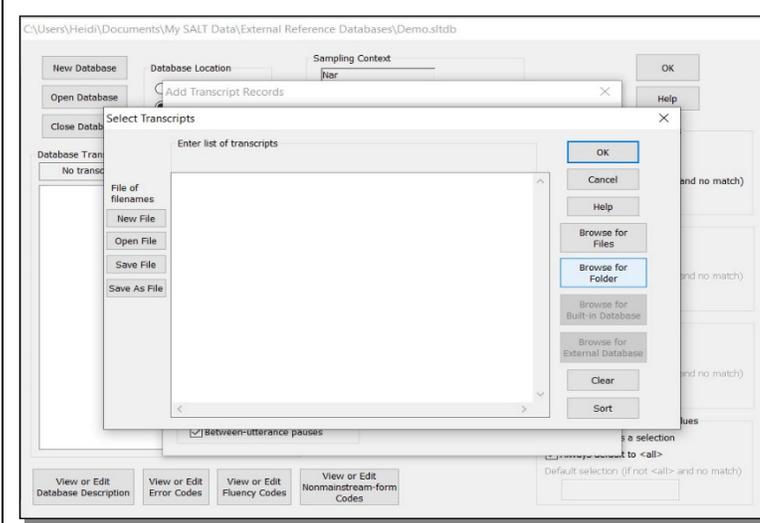
The next section specifies another default setting. When selecting samples from this database for comparison, users must select how samples should be equated for reports based on the length of the sample. I'll going to keep this default as "Words" so transcripts will be equated, by default, to the same number of total words.

We'll add the transcripts now and talk about the rest of the settings after that.

Select "Add Transcripts".

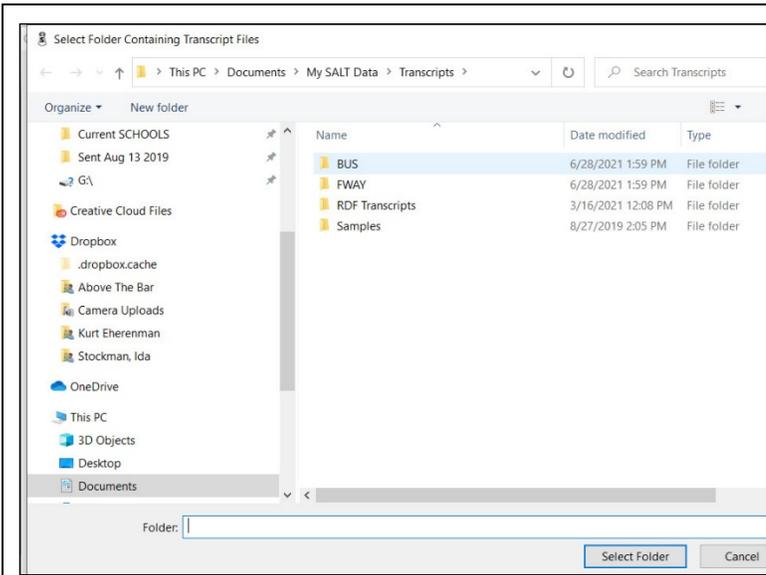


The "Select Transcripts" button is used to select the transcripts.

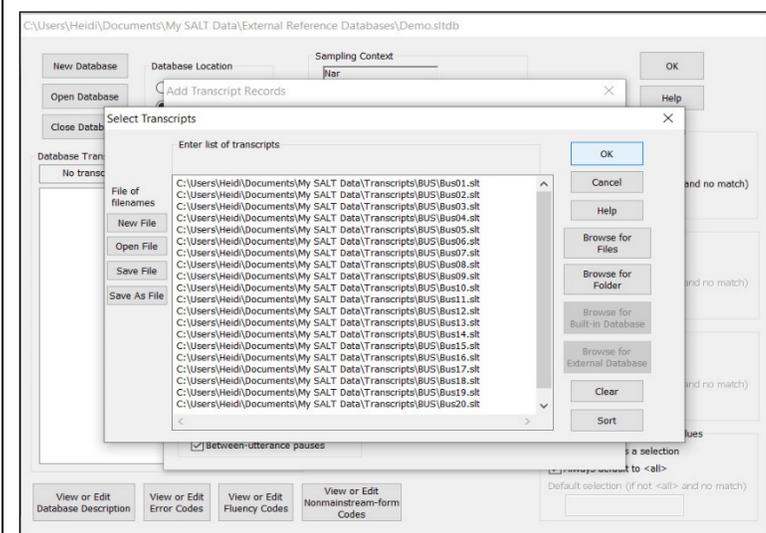


The transcripts for this database are story-retells elicited from young children. There are 40 transcripts in all. Half of them used the BUS story and half used Frog, Where Are You?. They will be added separately and, for convenience, are stored in separate folders.

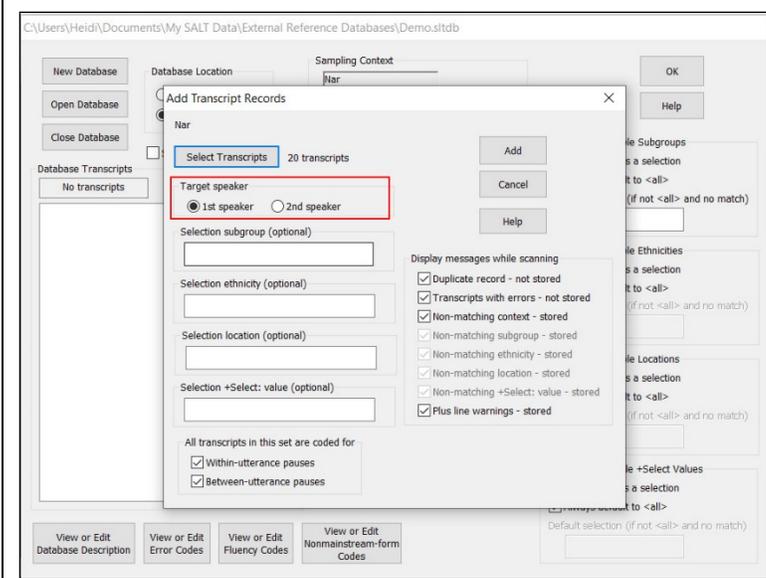
I'll browse for the folder.



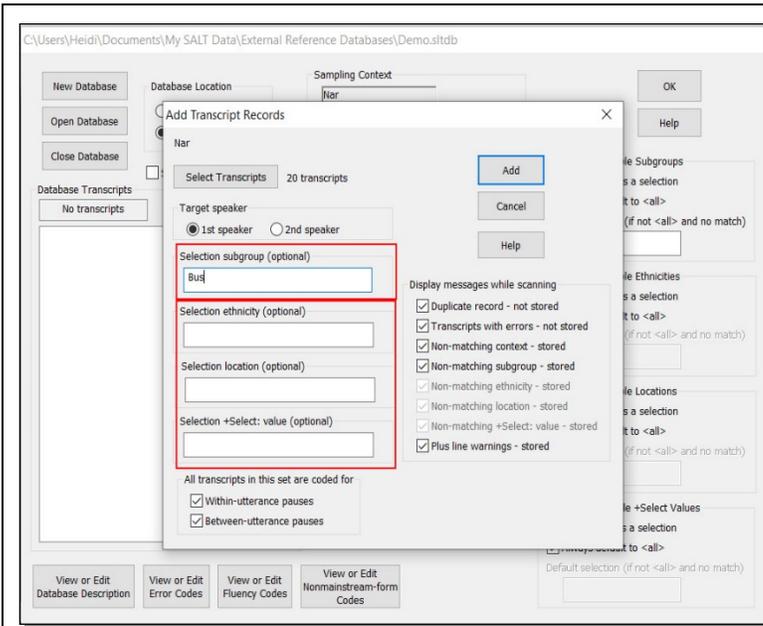
And go to where I stored the folders containing the transcripts. I'm going to add the BUS transcripts first so I select this folder.



The 20 BUS transcripts I stored in this folder are selected. Click OK.

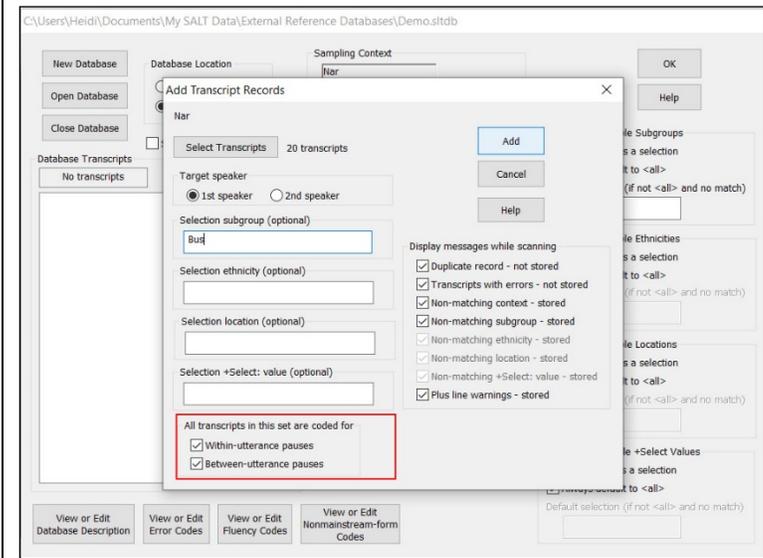


The target messages speaker is the 1st speaker.

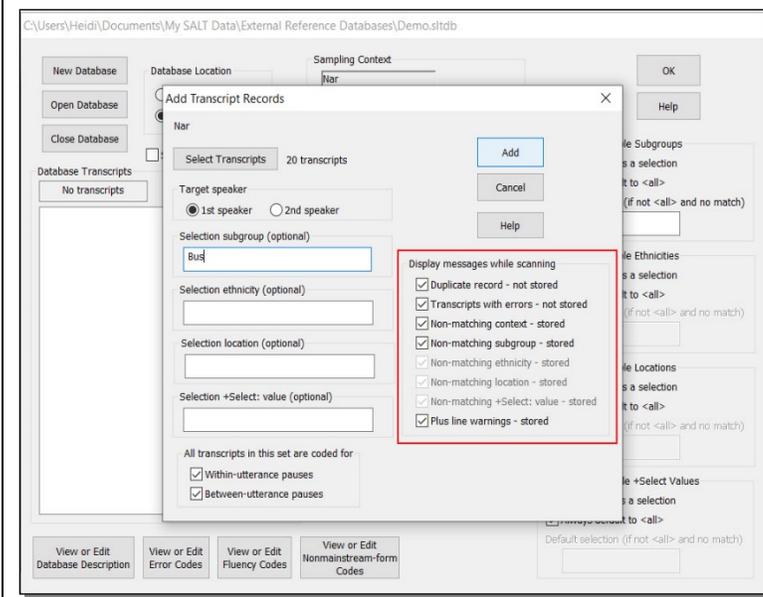


I'll enter "Bus" as the selection subgroup for this set of transcripts.

Notice that there are options for entering ethnicity, location, and +Select values. If you were defining groups based on ethnicity, you would enter the ethnicity for this group here. Similarly, if you were defining groups based on location or +Select values, you would enter the specific value for this group here.

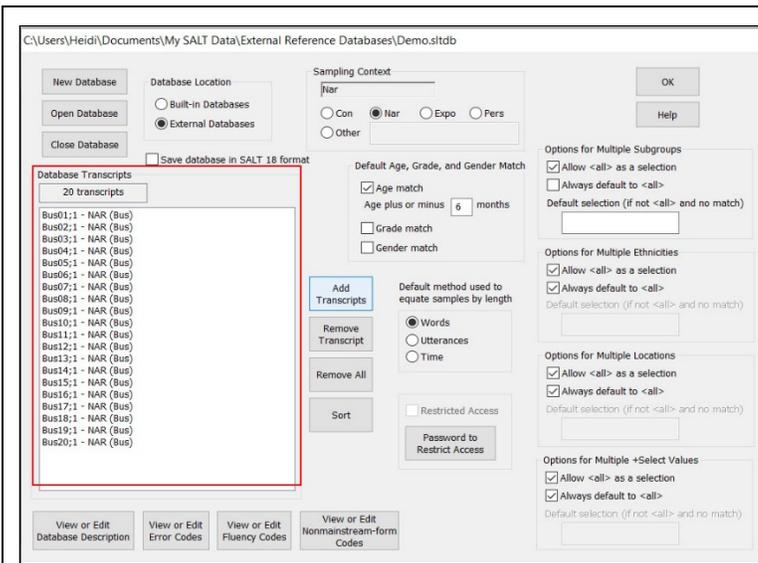


Notice the setting in the bottom left corner. This setting lets the database builder know whether or not this set of transcripts has been coded for pauses. When SALT is using this database for comparison, it needs to know which, if any, of the matching transcripts have been coded for pauses. Those not coded will not be included in any comparison pause data. And, unfortunately, SALT can't determine this by looking at the contents of the transcripts because the absence of pauses may either be because pauses weren't marked or because there weren't any significant pauses. When these transcripts are added to the database, they are flagged as either coded, or not coded, for pauses.

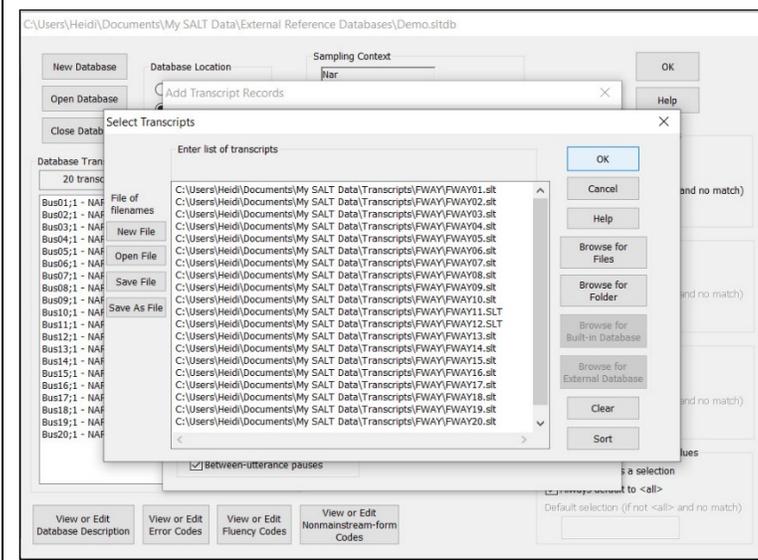


Now look at the messages available when these transcripts are added to the database. I like to keep them checked. Suppose, however, that these transcripts did not contain the subgroup plus line identifying the Bus story. You would then uncheck the option labeled "Non-matching subgroup - stored" to suppress this message. The transcripts would still be stored in the database as Bus story transcripts.

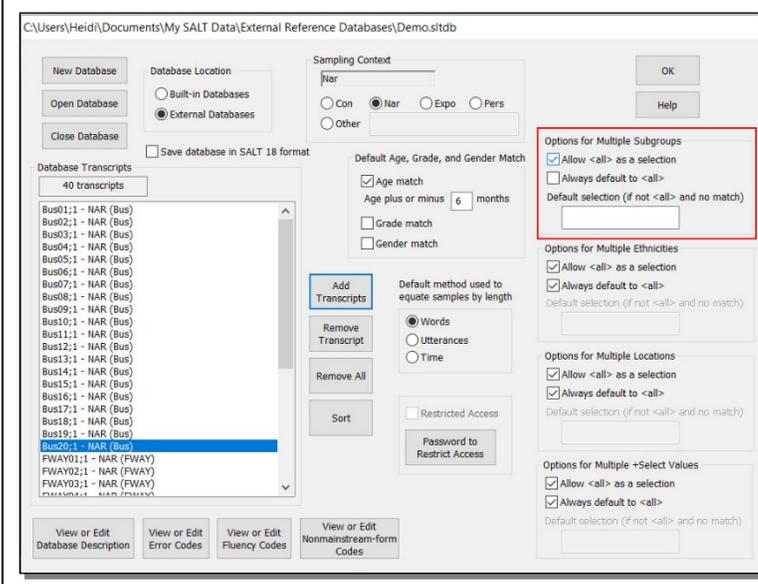
Add the transcripts.



The 20 transcripts are added. Notice that each transcript is identified by the name of the transcript, the context, and the subgroup.



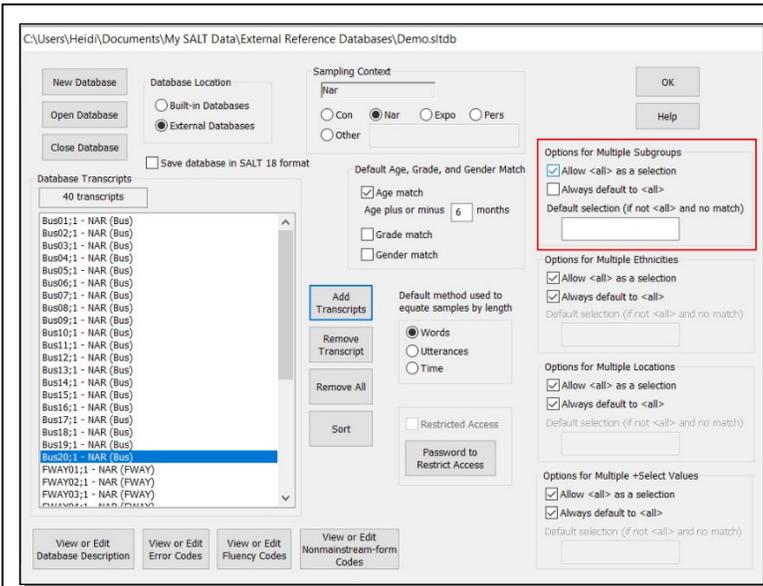
I'm going to quickly go through this same process to add the 20 Frog, Where Are You? transcripts.



Now let's talk about the rest of the settings.

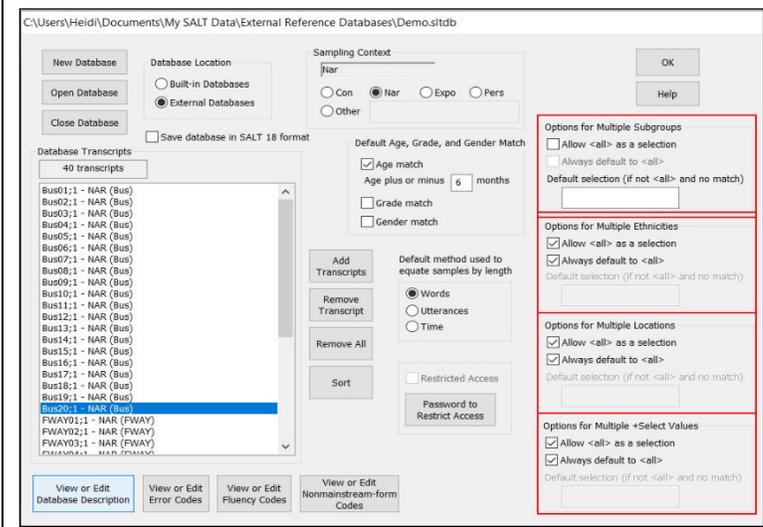
The "Options for Multiple Subgroups" setting is important because this database was built with 2 subgroups: Bus story and Frog, Where Are You?.

The first checkbox, "Allow <all> as a selection", determines whether or not **all** subgroups, i.e., both stories, is also an option. If you want users to have to choose between the Bus story and Frog, Where Are You? story, you would uncheck this box. If you want users to be able to compare their transcripts regardless of the story, you would keep this box checked. I want the users to have to choose their story, so I'll uncheck the box.

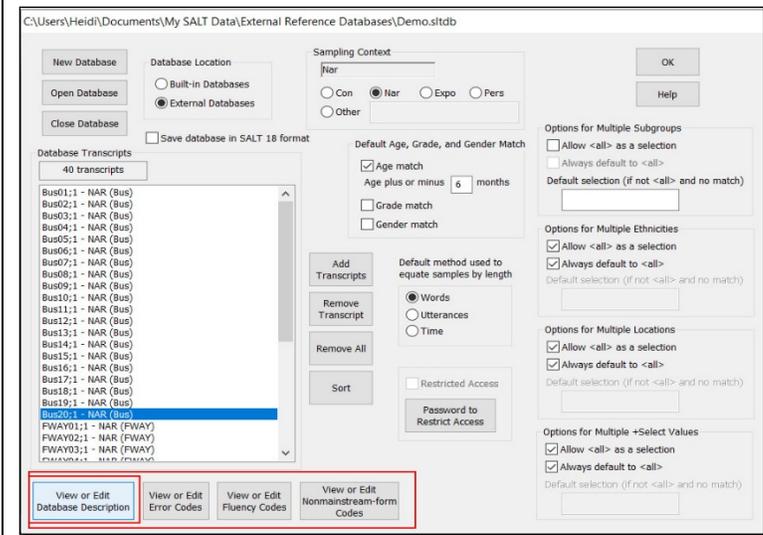


The next checkbox, “Always default to <all>”, is only available if the first box is checked. If <all> is allowed as a selection and this box is checked, <all> is always the default selection.

The text box labeled “Default subgroup (if not <all> and no match)” allows you to preselect one of the subgroups if the user’s transcript does not contain a matching subgroup. I’m going to leave it blank.

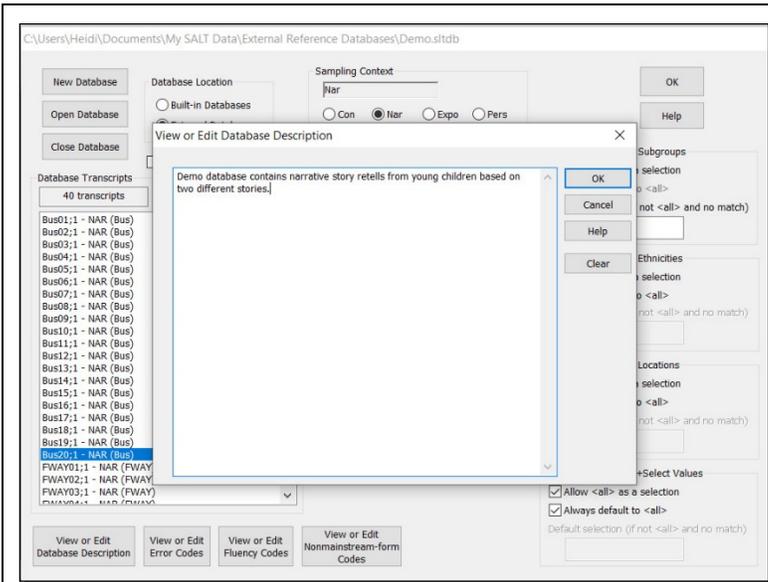


The three similar options function the same way if we had grouped the transcripts by ethnicity, location, or some user-defined selection values. I’m going to ignore them since they don’t apply to this database.

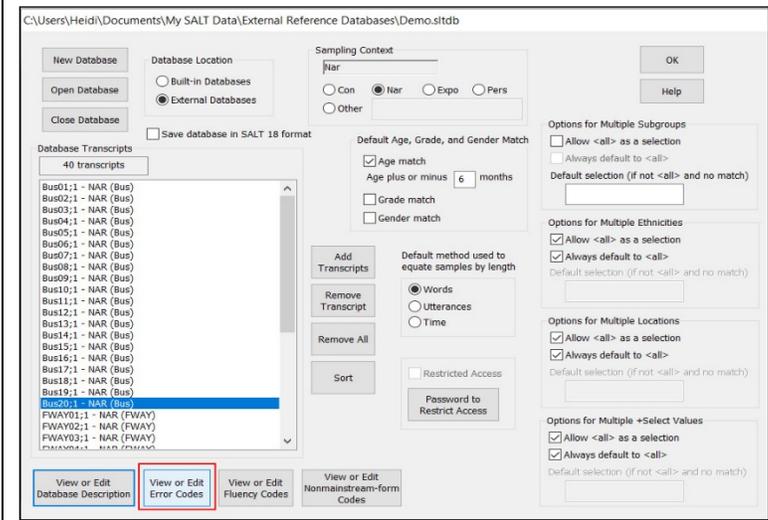


Now let’s look at the 4 buttons on the bottom.

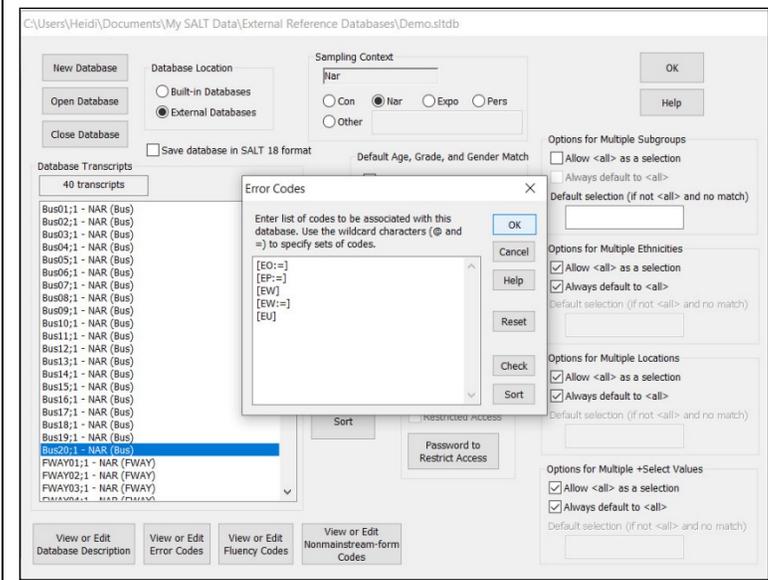
The first one, “View or Edit Database Description”, contains the database snapshot.



Every SALT database includes a short description, called the “Database Snapshot”, which users can view when selecting the database to use for comparison. That description, which is entered here, describes the participants, the elicitation protocol, any specified groups of transcripts, nonstandard transcription conventions, special coding, and acknowledgements. Although you can type this description in SALT when creating the database, you may find it more convenient to type it ahead of time using some other editor, and then copy and paste it here. I didn’t do that so I’m just going to type a few words here. And click OK.

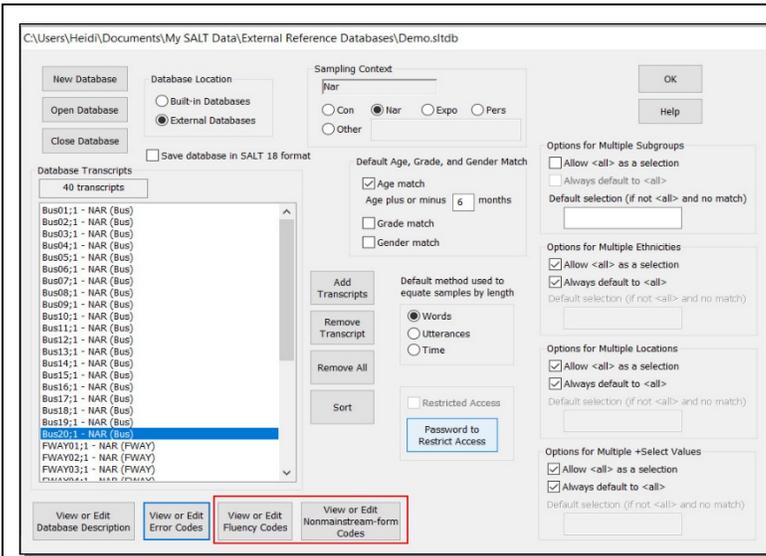


The next button on the bottom, “View or Edit Error Codes”, lets you define the error codes used in the transcripts added to this database.

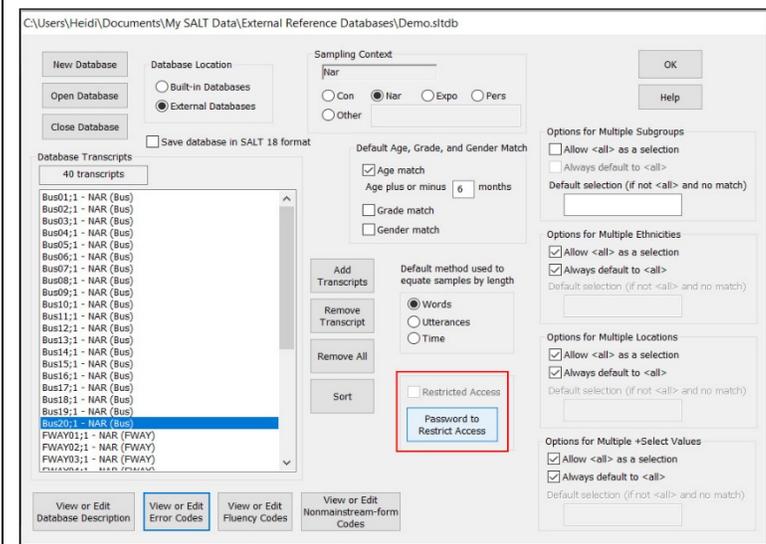


The list starts out with the default error codes. It may be, however, that different error codes were used when these samples were transcribed. If so, you would edit this list of error codes to match the error codes used. When comparing transcripts to this database, this list of codes defines the error codes for the database – but not for the user’s transcript.

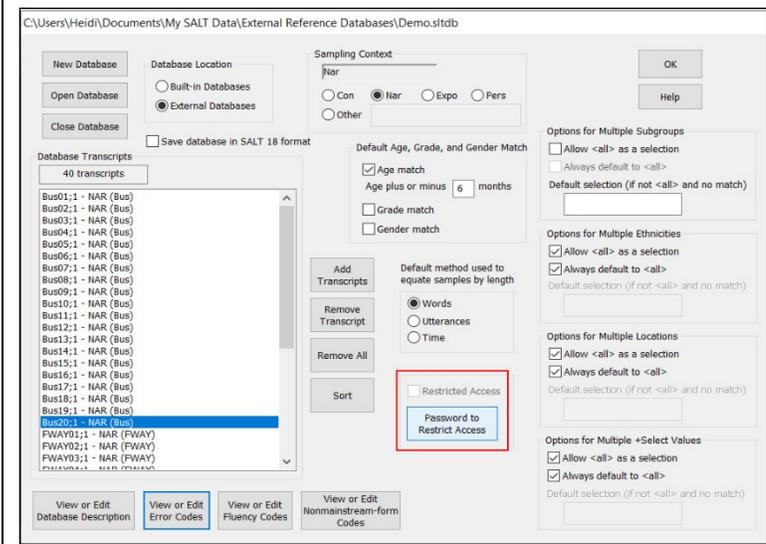
I’m not going to change this code list so I’ll just click OK or Cancel.



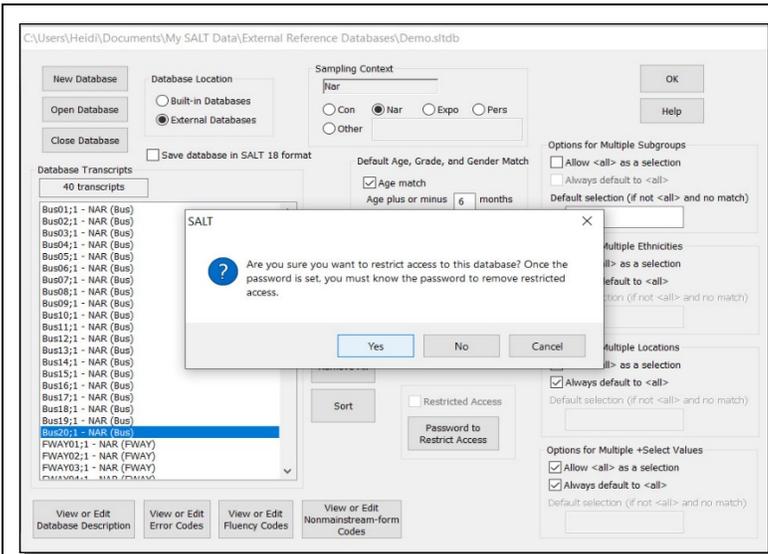
The other two buttons are used to change the list of fluency codes and nonmainstream-form codes used in this database. The transcripts were not coded for either fluency or nonmainstream forms so I don't have to edit them. If you were unfamiliar with these code lists, it might be a good idea to look at them on the off chance that these same codes were used for a different purpose.



Notice the button labeled "Password to Restrict Access". First some background... anyone using the "Rectangular Data File" research tool can access data from any SALT database on their computer. It is important to note that, although users can access the data generated from the transcripts, they cannot access the actual transcripts or any of the utterances in the transcript.



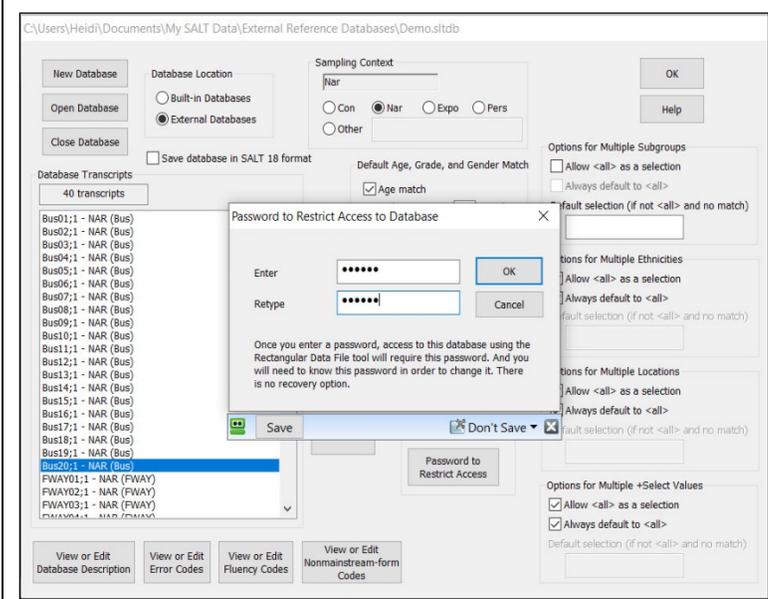
Now suppose I wanted to make this database freely available for anyone to use for comparison but I didn't want them to be able to extract data from it using the "Rectangular Data File". I would use this option to assign a password to the database. This password would not prevent anyone from using the database for comparison but if they tried to use it in the "Rectangular Data File" tool, they would be prompted for the password.



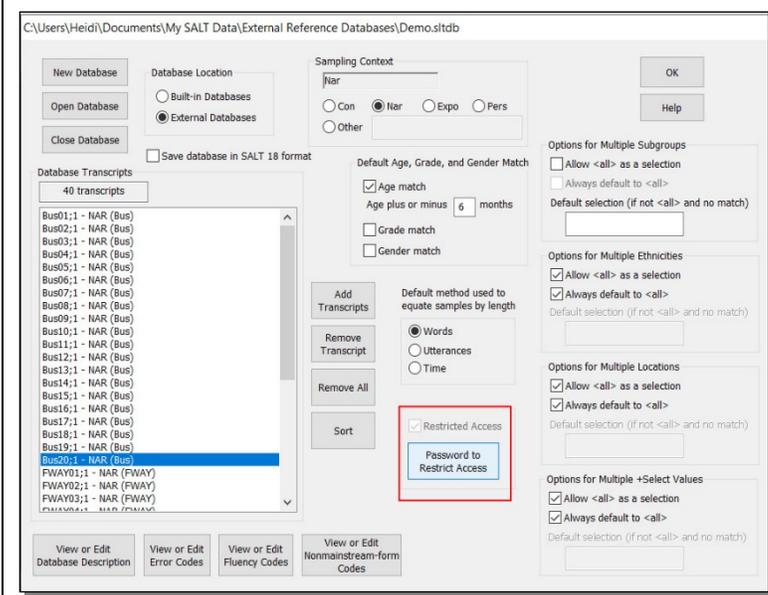
When I select this option, I'm warned that I must remember the password I set. If not, I would have to create a new database (which wouldn't be all that difficult if I still have the transcripts).

I'm going to go ahead and create a password for this database and then I'll remove it.

So, I'll click "Yes"

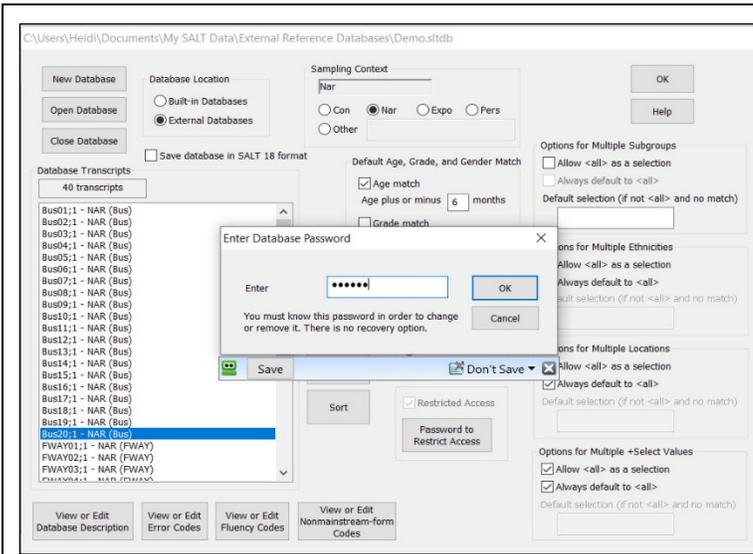


And then enter a password, retype it, and click OK.

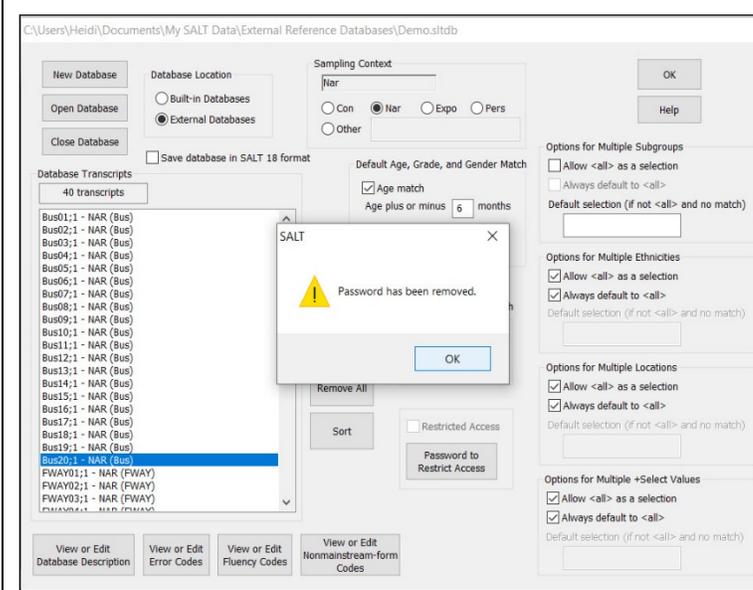


"Restricted Access" is checked.

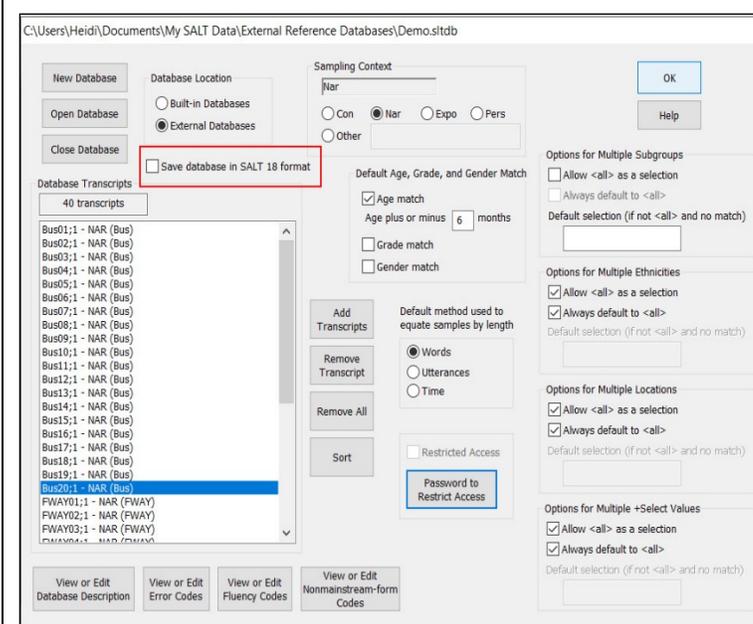
To remove the password, I'll click the password button again.



Then type in the database password, and click OK.



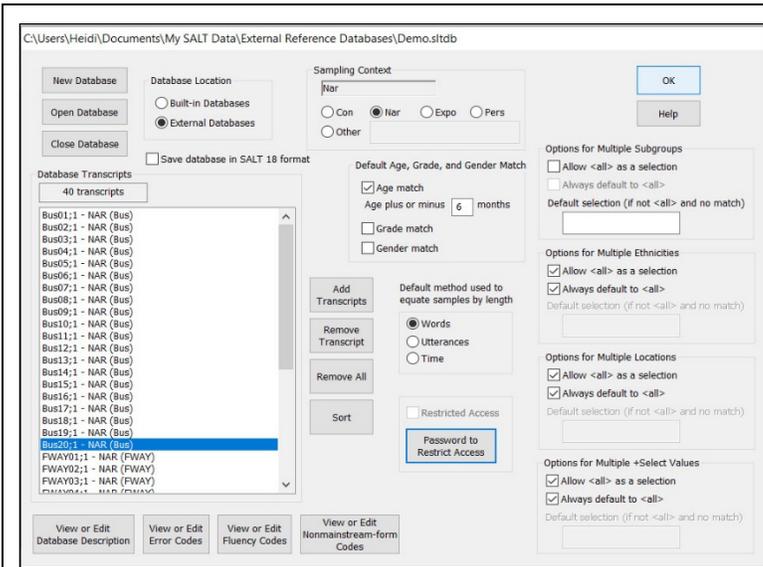
The database password has been removed.



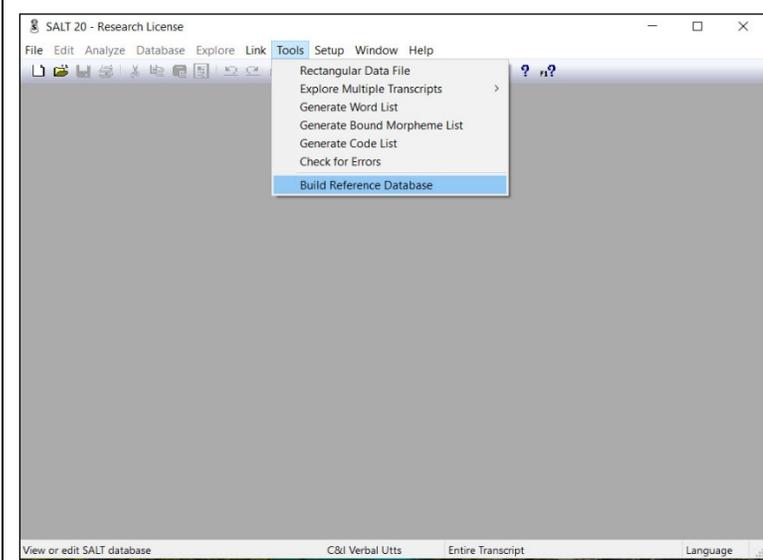
There is one last setting to talk about. Notice the checkbox labeled “Save database in SALT 18 format”. Several options were added in SALT 20 resulting in a slightly different database format which is not compatible with SALT 18.

The +Select: option was added in SALT 20 to allow an additional, user-defined, selection group. And the “Always default to <all>” options were added to get more control over the default selections.

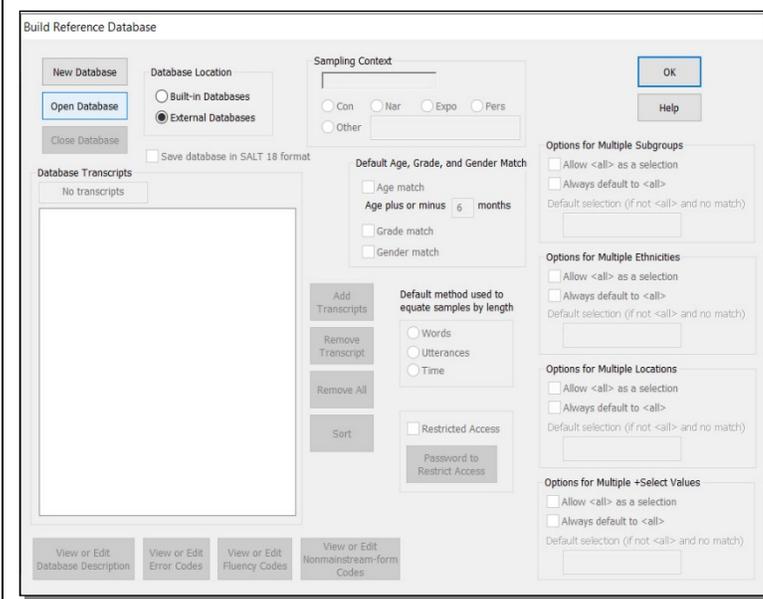
If you aren’t using these options, you may choose to save your database in SALT 18 format so that it is compatible with both SALT 18 and SALT 20.



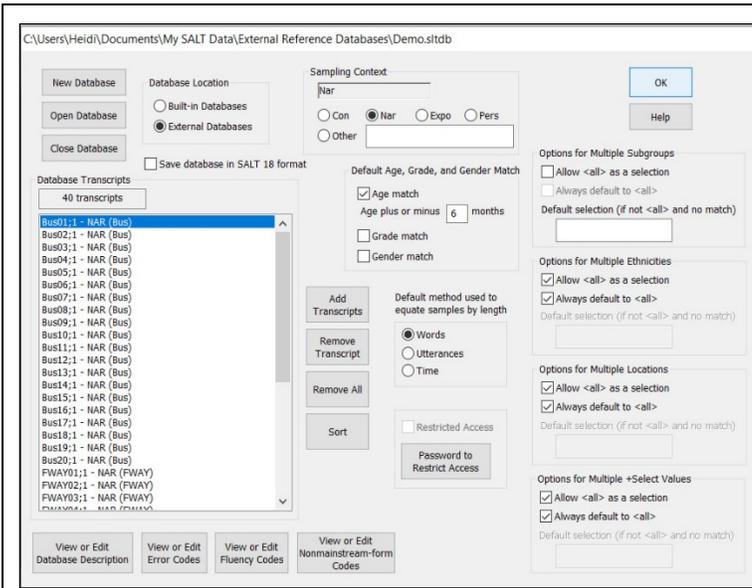
That's it. The database is created. The database is continually updated as you make changes. You don't need to explicitly save it. I can close the database or just close the dialogue box



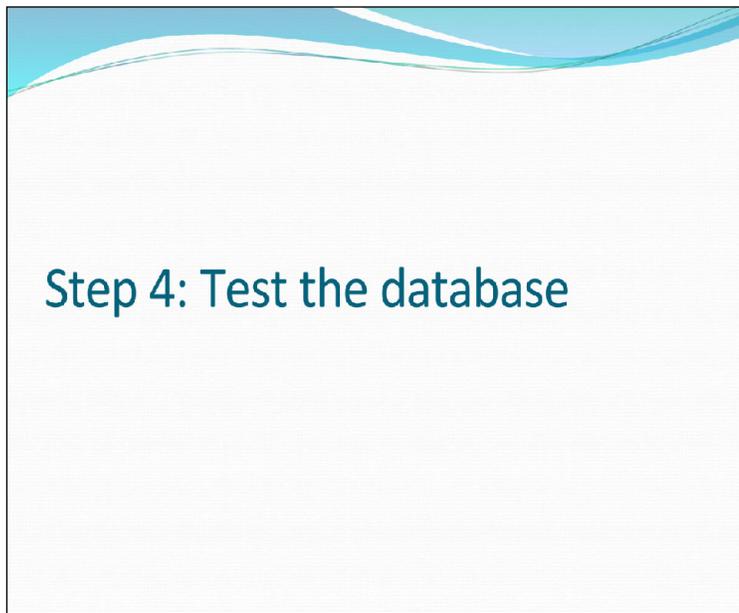
If I want to make changes to this database, perhaps adding more transcripts or editing the database description, I just select "Build Reference Database" from the Tools menu again.



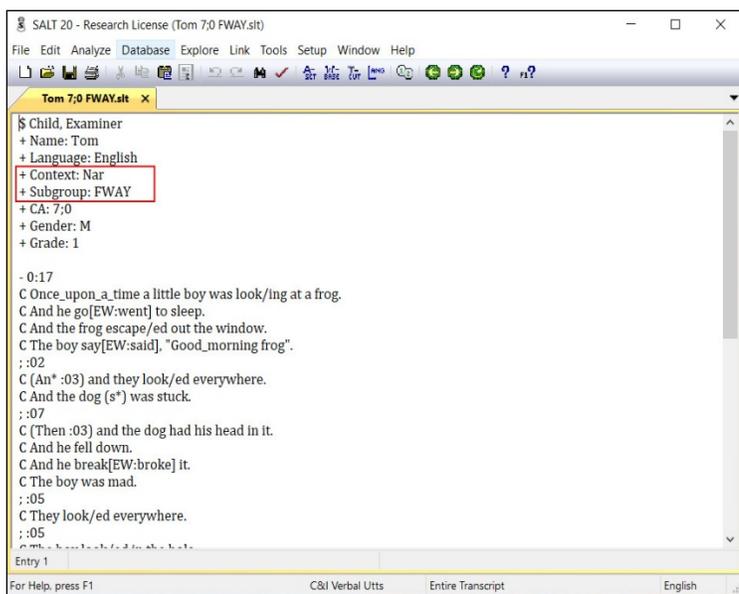
And select "Open Database".



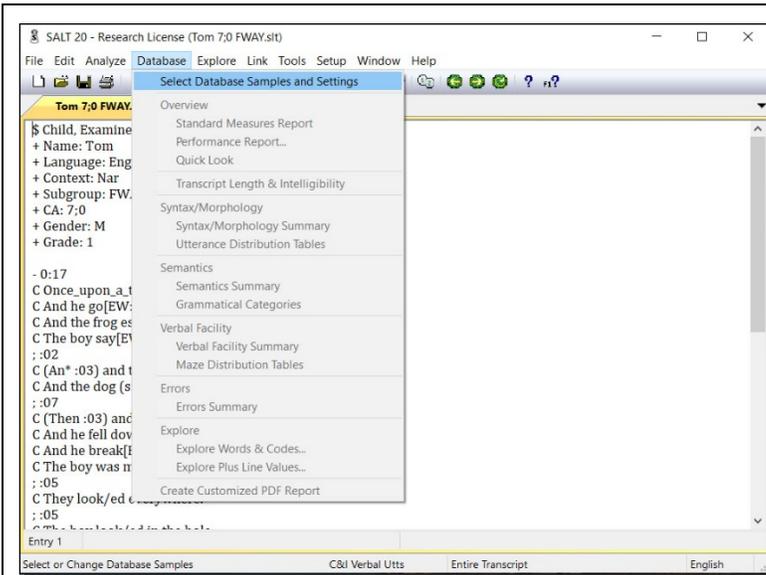
I'm finished with the database for not, so I'll just close the dialogue box.



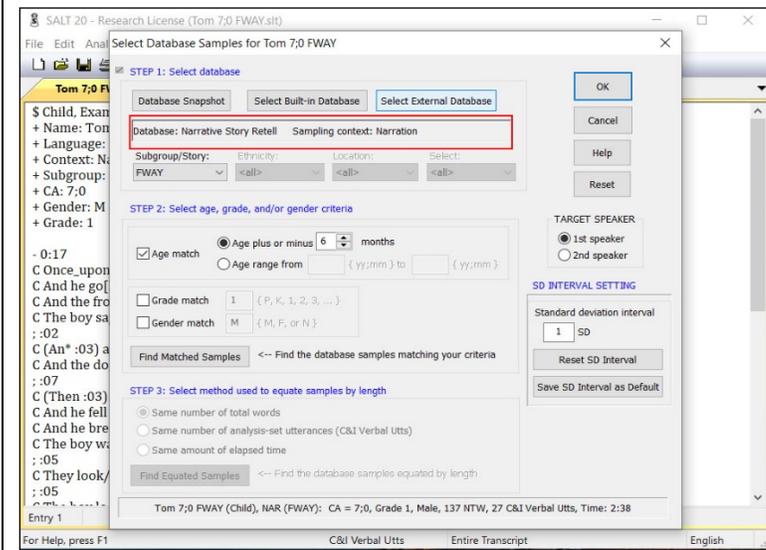
Now let's test the database.



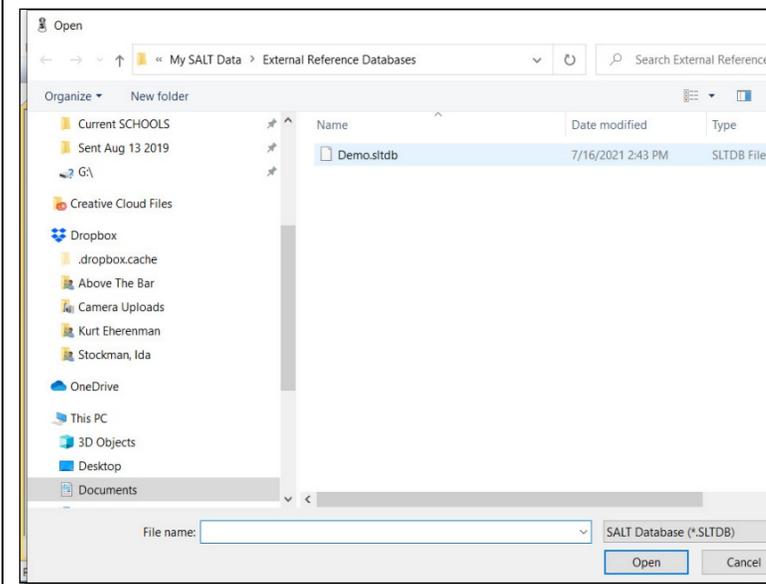
To test the database, I'll open a transcript from Tom retelling Frog, Where Are You?. Notice the context is Nar and the subgroup is FWAY.



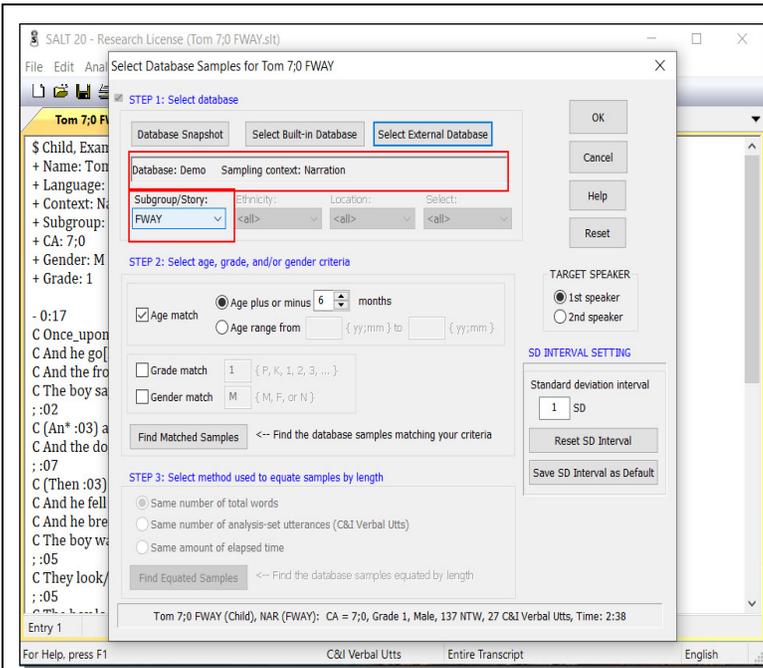
I'll go to the Database menu and "Select Database Samples and Settings".



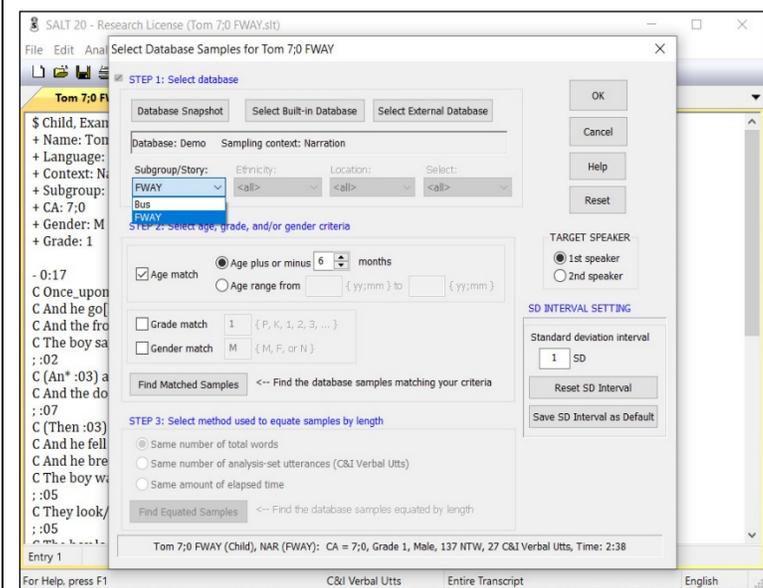
Notice the Narrative Story Retell database is preselected. To change this to the Demo database, just click "Select External Database"



And select the Demo database.

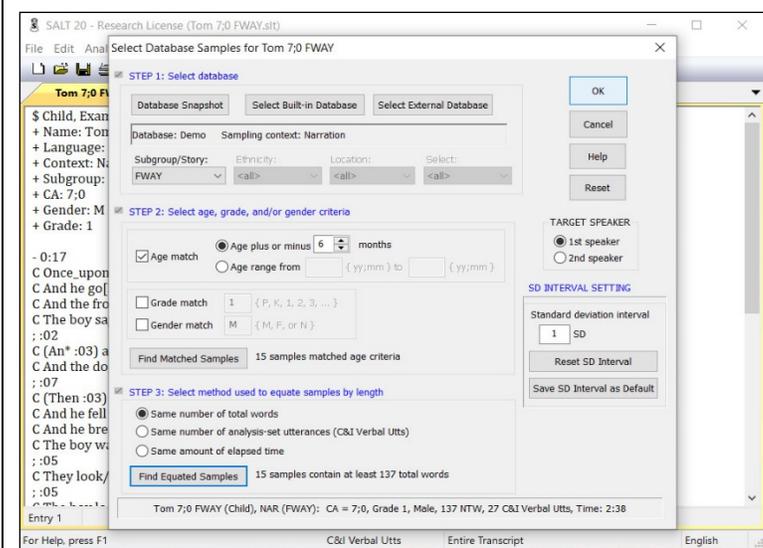


Now I'm using the Demo database for comparison. Notice the subgroup is preselected as FWAY because Tom's transcript specifies the subgroup as FWAY.



Now I'm using the Demo database for comparison. Notice the subgroup is preselected as FWAY because Tom's transcript specifies the subgroup as FWAY.

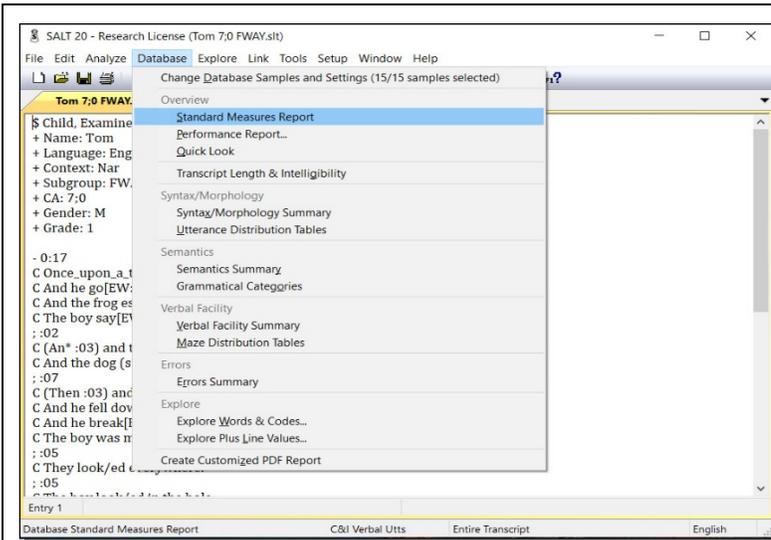
But if I click the dropdown arrow, you can see that the database also contains the Bus story.



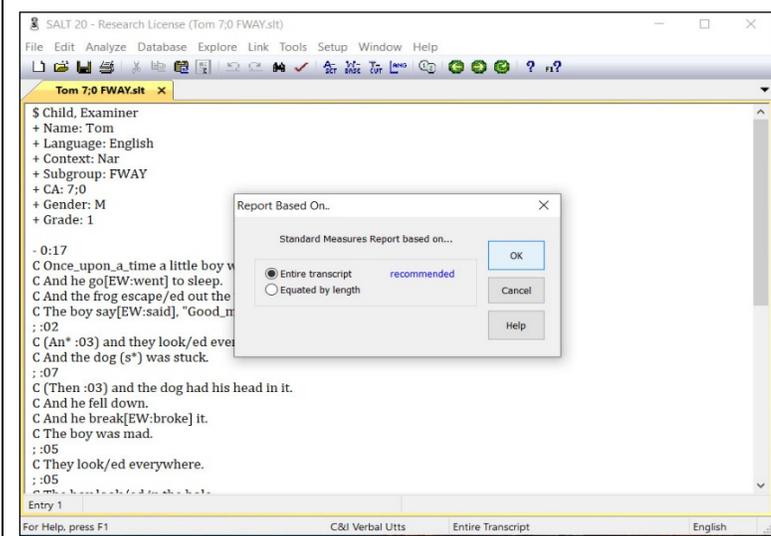
Keeping the subgroup as FWAY, I find the matched samples...

and the equated samples ...

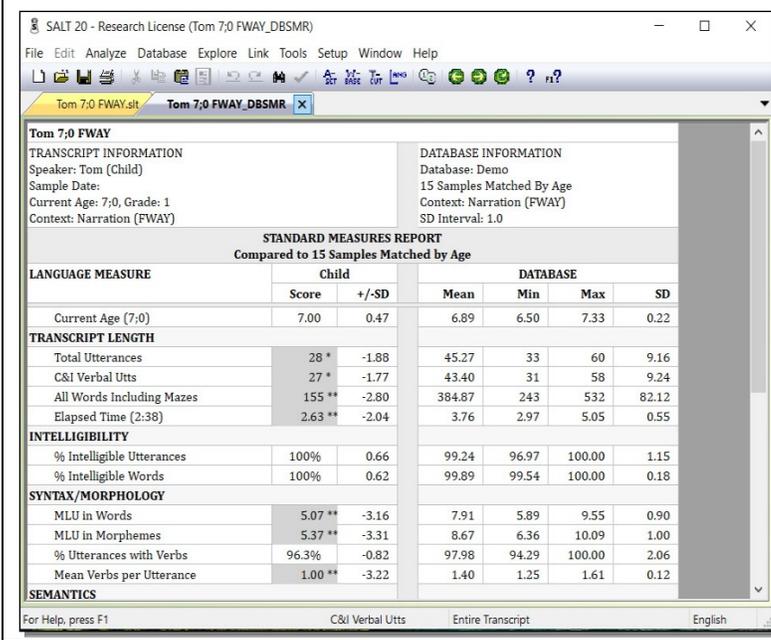
and click OK.



Now I'll select "Standard Measures Report" from the Database menu.



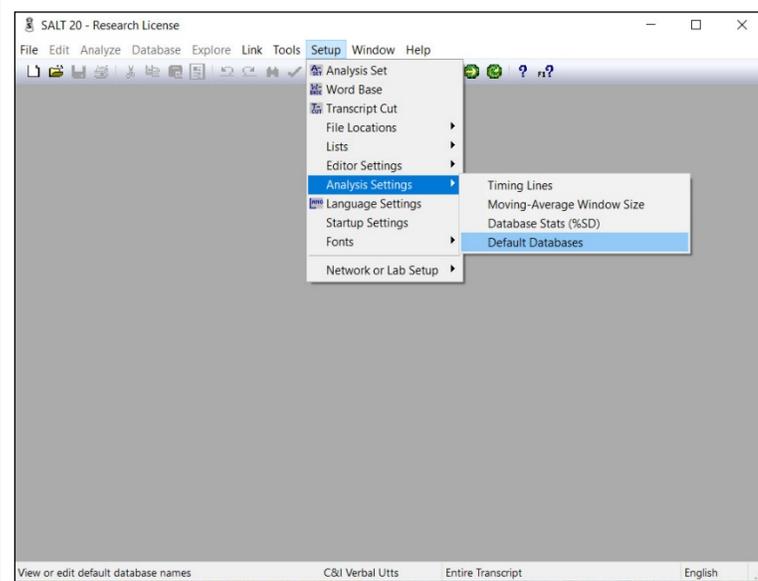
Based on the entire transcript.



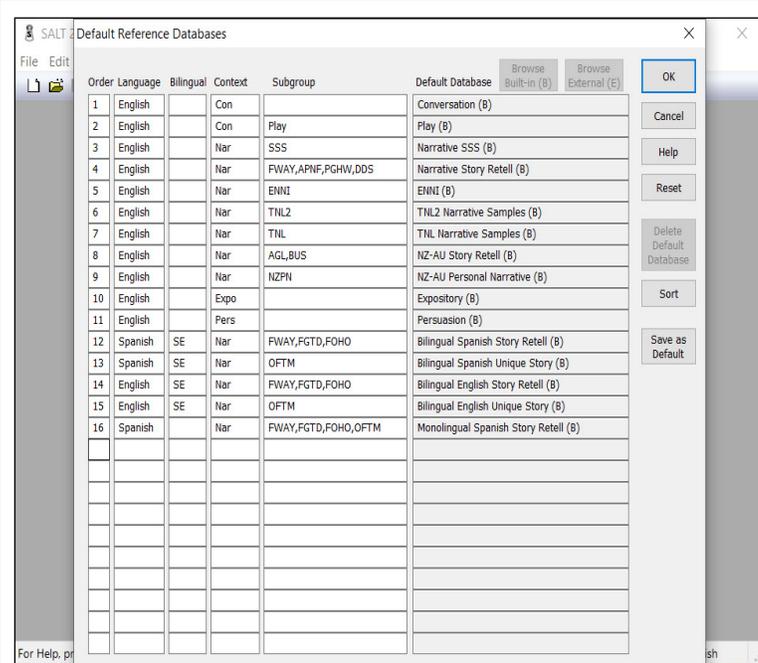
And the Standard Measures Report is displayed comparing Tom's transcript to transcripts selected from the Demo database.

Step 5: Edit the default database setting

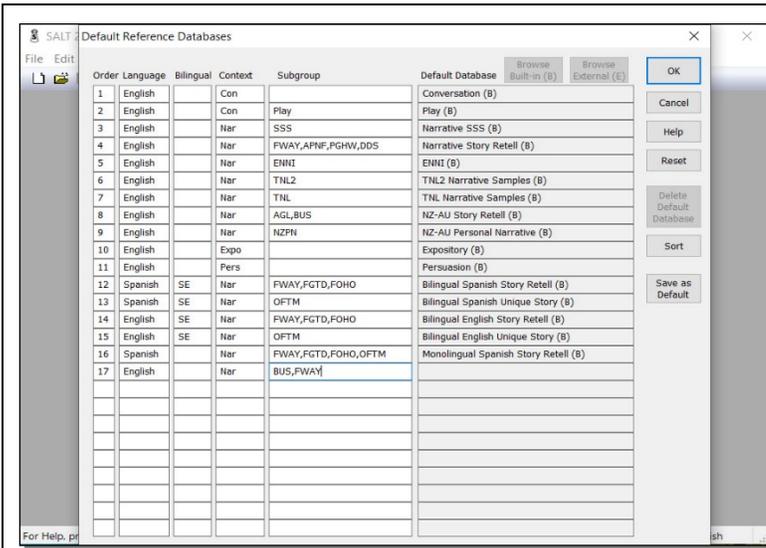
There's one final step which may make it easier to use this database. Remember when I compared Tom's transcript, SALT preselected the Narrative Story Retell database and I had to change it to the Demo database? SALT preselects the "best" database by looking at the contents of the plus lines and applying some rules. If I want the Demo database to be preselected, I need to change the rules.



Select "Setup → Analysis Settings → Default Databases".



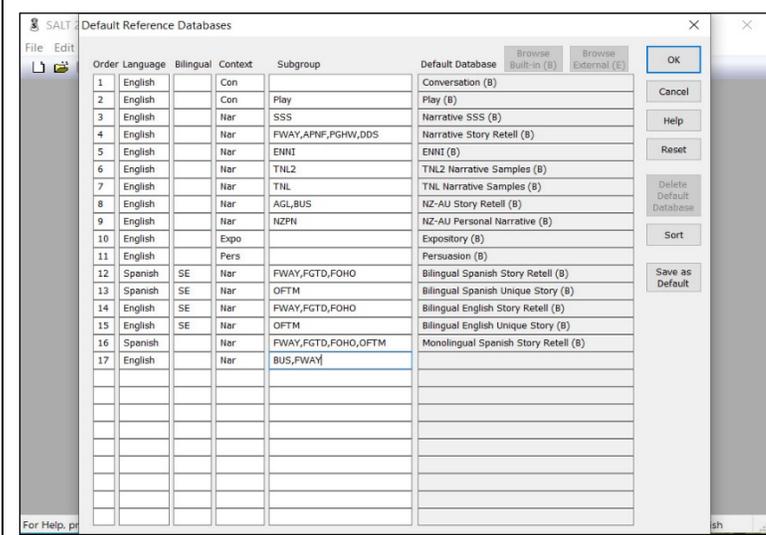
This dialogue box defines the rules for preselecting databases for comparison. The rules are based on language, context, and subgroup. If I want my new database to be preselected, I need to **carefully** add it to the list.



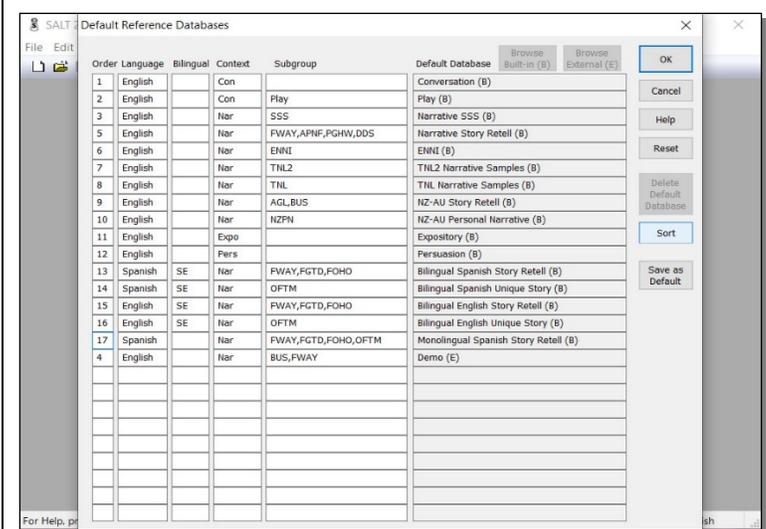
I'll start by adding it to the end as line 17. The Language is "English", the Context is "Nar", and the Subgroup is "BUS" and "FWAY". I'll click "Browse External (E)" since this is an external database and select the Demo database.

Notice line 4. The rule for the Narrative Story Retell database is very similar: English language, Nar context, and a subgroup which includes FWAY.

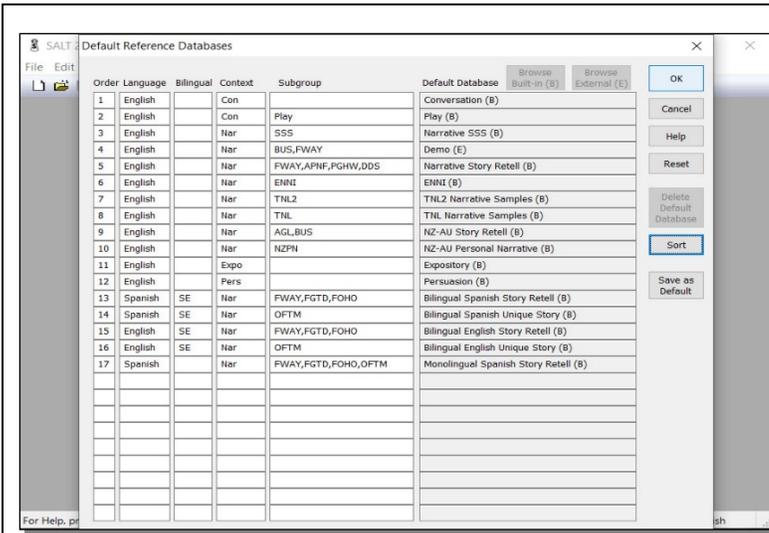
Also notice line 8 which is the rule for the NZ-AU Story Retell database. This is also very similar: English language, Nar context, and a subgroup which contains BUS.



When lines contain the same information, they are processed in order from 1 to the end. So I need to decide if I want my database to take precedence over one or both of the other databases. In this case, I want my database to have the highest priority so it needs to come before the Narrative Story Retell database on line 4.



So I'm going to renumber lines 4 – 16 to 5 – 17 and change the number of my database from 17 to 4.



Then I'll click the "Sort" button to arrange the databases in order.

Now, if a user compares a narrative transcript to database transcripts and the transcript subgroup is BUS or FWAY, the Demo database will be preselected.

If this is what I want, I would click "Save as Default".

Since this is just a demo database, I don't want to save it as the default, so I'll just click OK.



This concludes the tutorial on how to build your own SALT reference database.