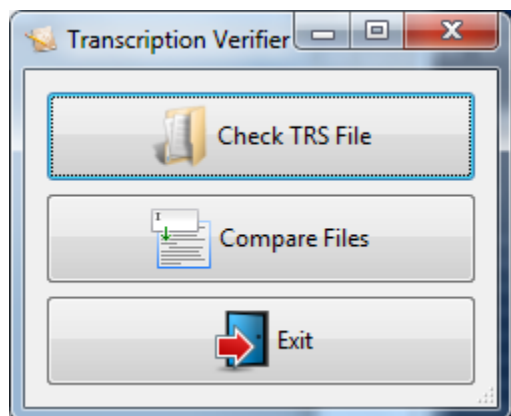# Transcription Verifier

Overview

This script checks for errors in TRS (transcriber) files. You can also use it to display the differences between two files.
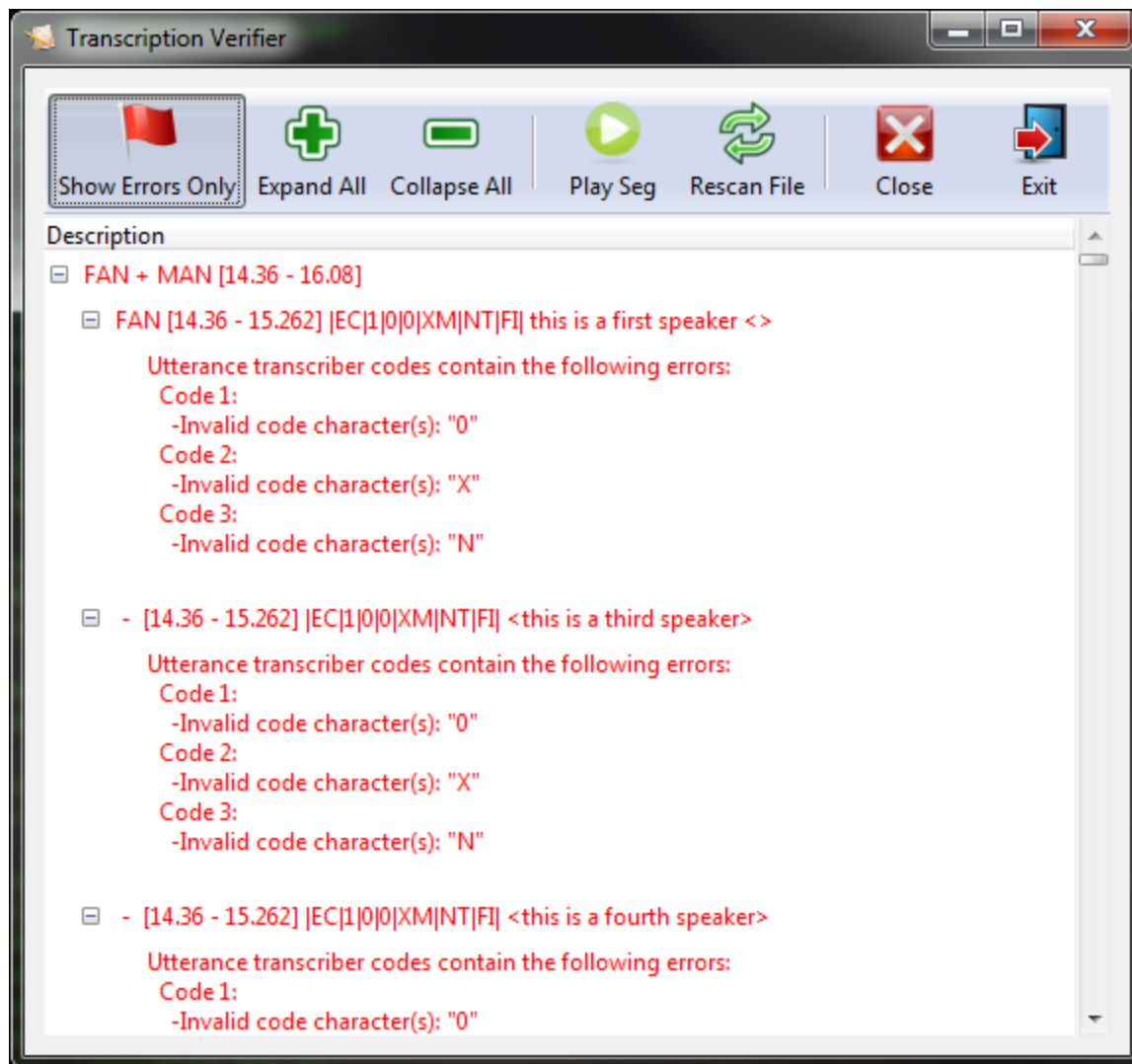
The main window looks like this:

I. Checking files for errors

To check a file for errors, click the "Check TRS File" button. A window will pop up asking you to select the TRS file you'd like to check. Select a file and hit "open". At this point, a progress bar is displayed. During this time, the program is looking through the TRS file, linking together segments joined with "I" and "C" codes, and checking for errors. This can take some time, depending on the number of segments in the file.

Viewing Errors

After the processing is complete, you'll see a window (like the picture below) showing the results. This window lets you view the transcriptions that have generated errors or warnings. If you like, you can also view the segments that have not caused any issues.

Note: If you like, you can open multiple TRS files at the same time (i.e. have multiple result windows open at once). To do this, just switch back to the main window and click the "Open TRS File" button again.

Transcription Verifier

Show Errors Only | Expand All | Collapse All | Play Seg | Rescan File | Close | Exit

Description

☐ FAN + MAN [14.36 - 16.08]

   ☐ FAN [14.36 - 15.262] |EC|1|0|0|XM|NT|FI| this is a first speaker <>

      Utterance transcriber codes contain the following errors:
       Code 1:
        -Invalid code character(s): "0"
       Code 2:
        -Invalid code character(s): "X"
       Code 3:
        -Invalid code character(s): "N"

   ☐ - [14.36 - 15.262] |EC|1|0|0|XM|NT|FI| <this is a third speaker>

      Utterance transcriber codes contain the following errors:
       Code 1:
        -Invalid code character(s): "0"
       Code 2:
        -Invalid code character(s): "X"
       Code 3:
        -Invalid code character(s): "N"

   ☐ - [14.36 - 15.262] |EC|1|0|0|XM|NT|FI| <this is a fourth speaker>

      Utterance transcriber codes contain the following errors:
       Code 1:
        -Invalid code character(s): "0"

The segments in the results window are displayed in a tree structure. The entries at the top level (the leftmost expandable bullet points) represent the segments, as they would be displayed in Transcriber. Each segment entry shows the speakers (with a "+" delimiting multiple speakers), followed by the start and end times in square brackets.

Each segment entry contains one or more transcriptions. In general, if the segment contains only a single speaker, there will be only one sub-entry displayed. However, there are several exceptions to this:

1. If a segment has an entry that contains the dot (.) operator (indicating that multiple utterances are contained within a single segment), the segment will contain two transcriptions. Both entries will display the same start and end time.

2. Similarly, if a segment contains multiple overlapping speakers (as indicated by the "overlapping" checkbox in transcriber), the segment entry will contain multiple transcriptions - one for each speaker.

3. If the segment contains multiple speakers that were indicated using angle brackets (<>) and/or newlines, the segment entry will contain a transcription for each additional speaker. In these cases, it is not always possible to determine who the additional speakers are - so, the description will contain a dash (–) in place of the speaker name (eg. FAN, MAN, etc.).

Colour-coding

Errors and warnings are listed below each transcription, like this:

- If an error is detected, its description is listed in red.
- If a warning is detected, its description is listed in yellow.
- If no errors or warnings are detected, the segment's description will be displayed in black.

 Button Functions

There are buttons at the top of the verification window. Here is what they do:

- "Show Errors Only" – This is a "toggle button" (It operates like a switch - when you click it once, it sticks down, and when you click it again it, it pops back up.) When it is "on", the list is filtered to show only those segments that contain errors and/or warnings. When it is "off", all segments in the TRS file are shown in the list (even those that don't have errors).
- "Expand All" – This expands all of the entries in the list.
- "Collapse All" – This collapses all of the entries in the list.
- "Play Seg" – This allows you to listen to a segment. If this is the first time you are clicking "Play Seg", a file selection window will open, prompting you for the corresponding WAV file.
- "Rescan File" – This causes the application to re-process the TRS file and update the results window. This may be useful if you wish to correct errors in the file using Transcriber while the transcription verifier is running.
- "Close" – This closes the verification window, but not the main application window. This allows you to select another TRS file without exiting the program completely.
- "Exit" – This exits the entire Transcription Verifier program, closing all windows.

Types of Errors and Warnings Detected

The transcription verifier checks for the following issues:

Errors:

- "Incorrect number of transcriber codes" – if less than 4 transcriber codes are found, an error is raised. Note that the application can't reliably differentiate between transcriber codes and LENA codes by letters alone (as some LENA codes use the same letters as transcriber codes). Instead,

the application identifies transcriber codes by their position (the last 4 codes on the line are assumed to be transcriber codes). *Therefore, if the transcription happens to contains 4 or more LENA codes, and no transcriber codes, the last 4 LENA codes will be assumed to be transcriber codes. In most cases, this will trip the invalid transcriber codes error (though it need not if the LENA code letters happen to be valid transcriber codes).*

- "Invalid code characters" – this error occurs when one or more of the transcriber codes contain letters that they shouldn't. The offending letters (and their positions) are listed in the error message.
- "Code contains more than one *X*" – This error can occur only for transcriber code 3 (since it's the only code that allows multiple letters), and it happens if there's more than one of the same letter present. Eg. "MCIM"
- "C without I, I without C, CI without I, CI without C" – These should be fairly self-explanatory. Links made by I and C codes will span segments that don't have any transcription. For example, you can link a FAN segment to a CHN segment if there are (untranscribed) silence segments in between them.
- "Ambiguous 'I's in segment." – This occurs when there are several segments in a row with 'I's (and no 'C's in between), and they're not differentiated with a numeral (eg. 'I1', 'I2'). A similar error can occur with ambiguous 'C's.
- "Codes containing U or F may not contain C or I" – This can occur only for transcriber code 3, according to the transcriber manual.
- "Unrecognized LENA Note" –This occurs when a LENA note (eg. "VOC", "CRY") is found that is not recognized.

Warnings:

- "Utterance has no transcription phrase" – If no transcription phrase is found, and the segment is not silence or noise, this warning is raised. It is only a warning because in some cases the speaker may not have said anything coherent and so there is nothing to transcribe (except the codes).

II. Comparing files

You can also use the program to highlight changes between two transcriber files. To do this, click the "Compare Files" button in the main window. You'll see a window that looks like this:

Enter the following information for each file:

- "Transcriber Name" - If you like, you can enter your name in this box – these names will be shown on either side of the split view (described below) so that you can tell who the work on each side belongs to.
- "Path" – this is the path to the TRS file you want to compare(you can use the "browse" button to select one).

Once you're done, hit "OK". At this point, the program will compare the files and open up a window that shows you the differences, like the one shown on the next page. This may take some time, depending on the length of the file.

The table shows one row for each segment whose transcriptions are different. In this example, the left side shows segments from Horace's file, and the right side from Gertrude's. The highlighted words are the differences. Each entry shows this information:

The highlighting colour shows the type of change that was detected, as described in the legend.

A couple of notes: If two segments were joined with the dot (.) symbol, they will appear separately, but with the same start and stop times. If a segment has multiple speakers (like the top one on Harace's side), the table row will contain multiple lines – one for each speaker.

Don't worry about the links and line numbers – they're purpose is to help you find the next segment in the event that you're showing the whole file (including the segments that match), but we're only showing the differences.