Oral History & Technology

WhisperCorrector manual

General

With support of the PDI-SSH project

WhisperCorrector software (current version: 1.4.2) is intended for correcting Automatic Speech Recognition (ASR) results in srt and/or json format, as by Whisper. The software converts the recognised text (the source file) into an xml file and displays the text next to the audio or video file. Currently, Whisper cannot yet recognise speakers or indicate speaker changes. In WhisperCorrector, you can add different speakers, correct the text, replace one or more words at once, add full stops and commas, run-out characters behind unfinished sentences, etc. Once finished, the corrected text can be exported all at once or separately to 3 different formats:

1) a subtitling format: SRT, can be placed under video files in video editing programmes
2) a simple text format: TXT, can be used for Forced Alignment*
3) a 'word.txt' format for word processing in MS-Word, with formatting so that the left margin can be set to indent 3.5 cm, putting the speakers in front of the text.

* At this time, 24 July 2023, the FA tool from the Centre for Language and Speech Technology at Radboud University has not yet been converted to Whisper. WhisperCorrector v.1.4.2 has already integrated a nice degree of alignment.

Whisper

With the use of the so-called End-to-End speech recogniser like Whisper, speech recognition has become quite good. Dutch, English, German, Italian and Spanish gives ±94% correct recognition of spoken words. So it is well worth using this form of ASR and then correcting the results if necessary. Manually transcribing 1 hour of spoken text takes around 6-8 hours. With the combination of ASR and correction, this drops to 1-2 hours, depending on the quality of the audio recording, clarity and accuracy of the spoken language, etc. In addition, what Whisper does is reasonably correct punctuation. Full stops, commas, question marks are placed properly and capital letters are mostly used where they belong.

A minor drawback of Whisper is that repetitions and unfinished sentences are not displayed. Whisper, coming from OpenAI, probably uses something from the Large Language Model (GPT3 or GPT4) for recognition and that ensures you get back mostly nice sentences.

So, you can correct the results with WhisperCorrector.

WhisperCorrector

WhisperCorrector is open source software, free to download from https://speechandtech.eu as a zip file (Windows) or a DMG file (Apple). Permission may need to be granted on the computer, under System Settings/Privacy and Security, to open the downloaded software. You can edit and save your material on your own computer in a secure environment of your choice.

Start WhisperCorrector

To start with, you will have to import the transcription result once. The file type, depending on the version of WhisperCorrector, is an srt or json file; both are fine. The first time you open WhisperCorrector, you will see a rather empty screen (fig. 1).

Open Settings once to select the right settings to import the recognised material. The audio/video is usually recognised if it is in the same folder, and the files all have the same name.

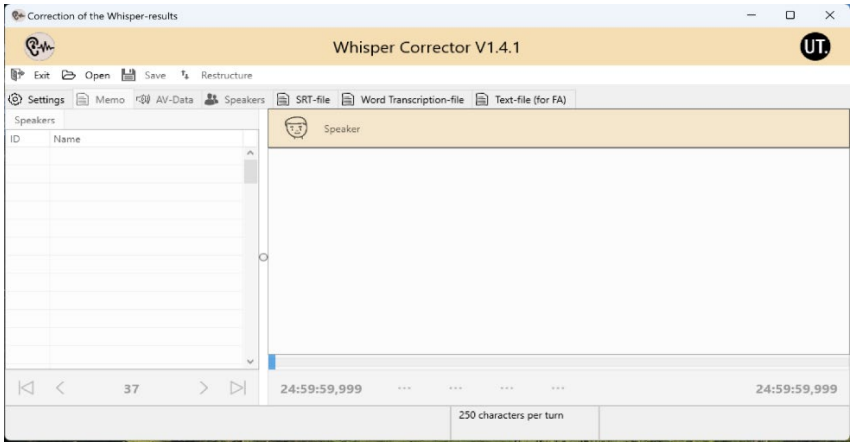

Fig. 1 : The image of the WhisperCorrector when you launch the software. The version shown (V 1.2.3) contains a mix of Dutch and English titles/headers. Once the software is "ready" we will change it to real English

To get started, you will have to import the transcription result once. The file type is either an srt or json file, depending on the version of the Whisper recogniser. Probably in the near future this will be json only.

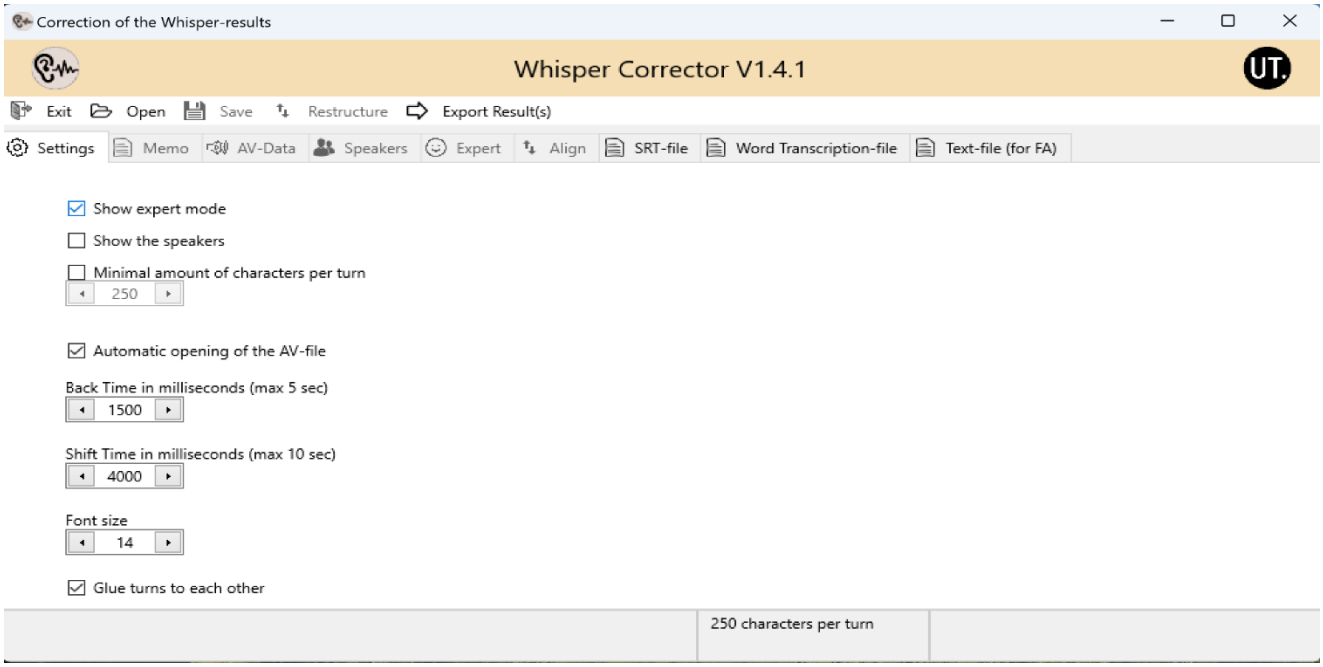

Fig. 2: Select Show Expert Mode for single read ASR results under 'Open', select Import SRT.

Optionally, fill Minimal amount of characters per turn for the amount of text in the memo view, with at least 250 characters or as many more as desired. If you leave it blank, each line of text will fill one 'turn'. If you read in a json file, 250 characters will be read in by default (Fig.3). Choose longer pieces, e.g. 500 or 1000 characters depending on your working method. You can also combine turns later when you don't hear transitions properly, for instance. The total number of turns is shown at the bottom of the mem-display (Fig.3), and can increase considerably when reading short pieces.

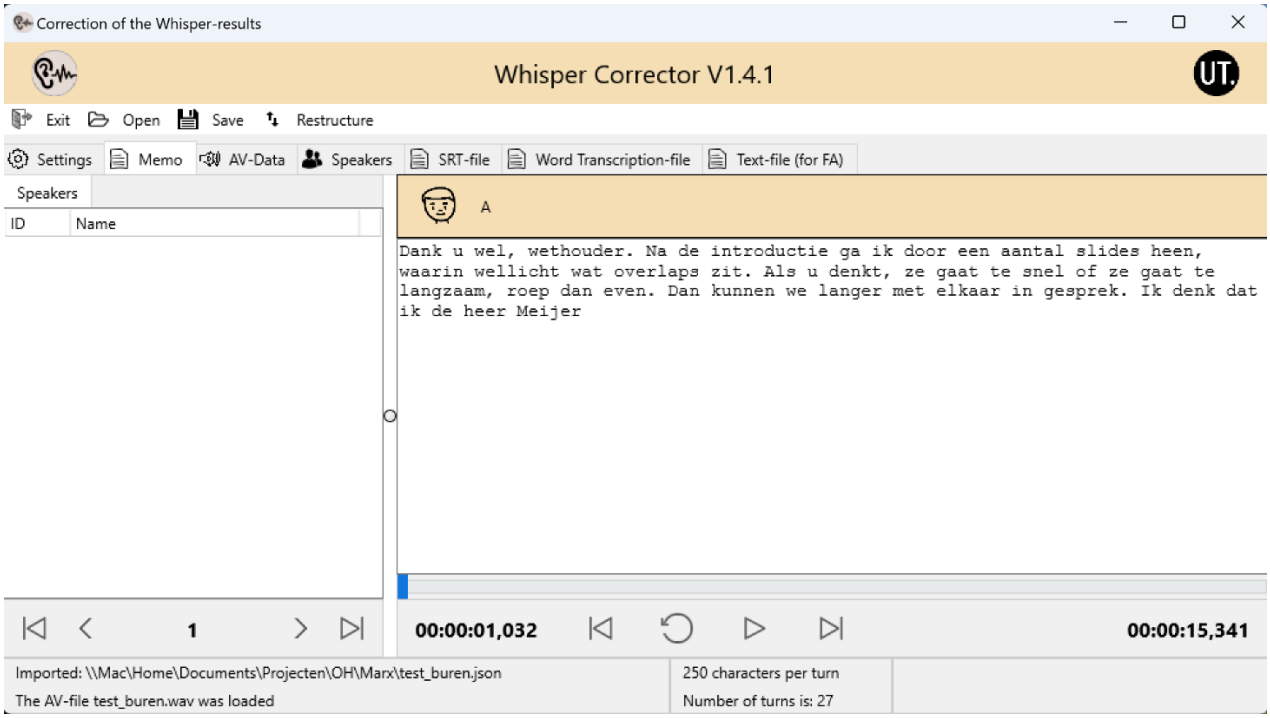

Fig. 3: Import a json or srt file. By default, a json file reads ±250 characters

'Show Expert Mode' becomes default after you read in a json or srt file. After reading, click 'Save' and then 'Save xml'. You won't need the srt or json file anymore now.

When you are done correcting, export the result. To do that, click Show Expert mode again in settings. Depending on whether you are working on a Windows or Apple machine, when exporting you may be asked to replace the original result with the corrected version (see Exporting results)

Add speakers

The last step is to add speakers. Click on the Speakers tab and fill in the details for each speaker. The ID field is generated automatically. The Gender, Role and Description fields do not affect the correction process or the output and can also be left blank. Role: Interviewee or Interviewer. The Name field is carried over into the transcript as filled in. It is important not to make these unnecessarily long and get them right the first time. Improving is possible, but if you are already far with correcting and the speaker is speaking a lot, it should be updated manually per turn. The old name is no longer recognised as the speaker and ends up in square brackets between the spoken text. The speaker who is speaking most often can be marked as default. This is usually the interviewee. This automatically appears as the speaker at the top of every turn (can be changed by right-clicking). All speakers completed? Click 'Save all Speakers' (fig.4).

Fig. 4: Addition of speakers.

Save XML file

Once all this is done, save the results as an XML file under 'Save'. You will always use the XML file from now on when correcting transcriptions.

Open XML file

After you have read the srt/json files and saved them as an xml file together with the audio/video file, the correction phase of speech recognition begins. From now on, you always open the xml file - the work file (fig. 5). You can always save and close WhisperCorrector in between and continue later. But... once you have done something, always click Save -> Save XML before closing WhisperCorrector.

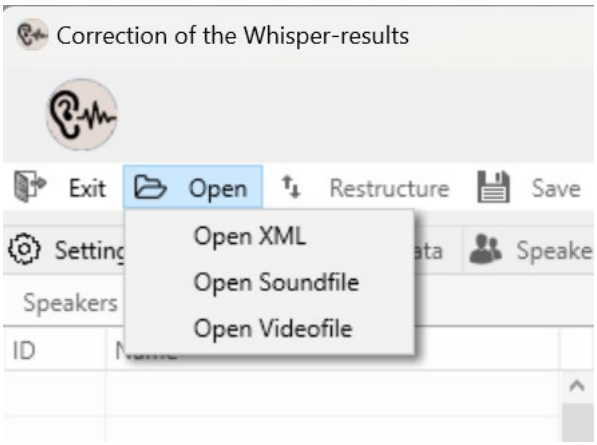

Fig. 5: Open XML to correct

Correction

Click on the Memo tab to start correcting the ASR results. On the right, you see the recognised text that you can correct like in a word processing programme (functions see below). Below it, audio buttons and the start and end time of this fragment. The 4 buttons in the middle are for playing/pausing the audio fragment. The bar just above it can be used to drag to a specific moment in the sound clip.

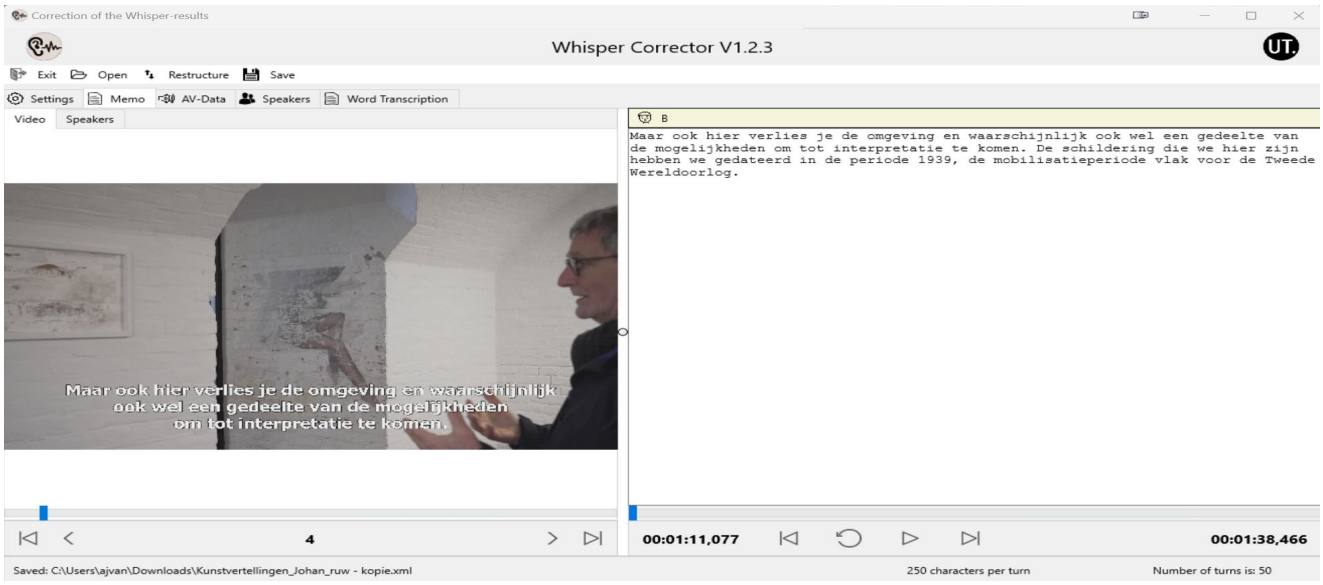

Fig. 6: Here you can hear and read what was said and recognised. In the text field, you can correct.

On the left panel you can set 2 tabs open, Video or Speakers. Below are buttons to scroll through the audio/video file -10 'turns' back, or in steps of -1 < or +1 > or +10 (fig. 6).

Insert speakers

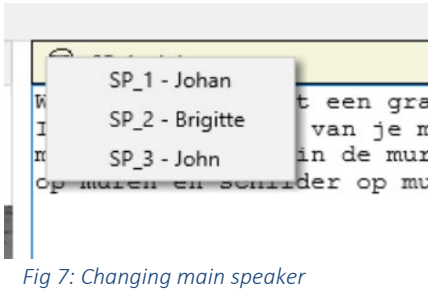

Fig 7: Changing main speaker

By default, the default speaker is at the top of each turn. You can change this on the spot by right-clicking on the head in the bar above the text field and selecting the appropriate speaker (fig. 6). However, it also often happens that you have a speaker change somewhere in the text field and therefore need to add a speaker who starts speaking. To do this, go to the place in the text where you want to add a new speaker. Right-click and select the top item (Add a speaker or Add 'n speaker). Then select the appropriate speaker again. The software adds the speaker with its ID and name, for example: [T3 - Peter Verboom], (fig.7). The hard returns have no meaning: they are only added to improve the readability of the text.

Fig. 8: Adding another speaker in the middle of the turn's text.

Combine turns

You may want to combine what you hear with the previous or next turn. This happens especially when the turns are quite small. To do so, right-click on the turn number in the bottom-right window and select **Glue turn with prev one** or **Glue turn with next one**. The turns are then glued together, and the text is merged into one. Before doing this, make sure you save your corrections, otherwise they will be lost in the gluing process.

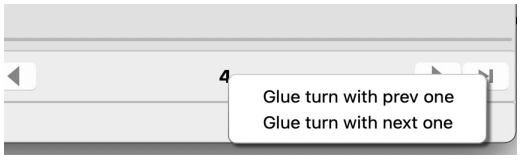

Fig. 9: Bonding of two adjacent turns

Find and replace words

Names that are not capitalised can be searched and replaced with another word or word combination. To do so, select the word(s) to be replaced, right-click on the selection, and select **Search & Replace**. The software copies the word into both the Search and Replace fields. When you click Ok, the original word is replaced with the new word throughout the document.

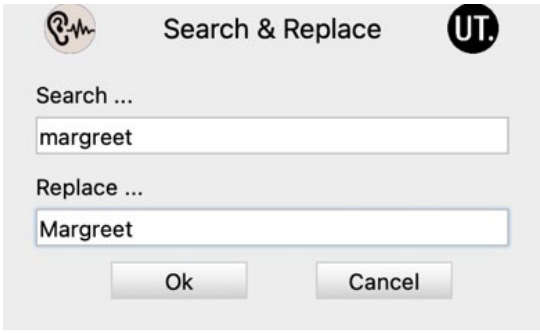

Fig. 10: Right-click function: Search and Replace all. Also useful to find and correct typo's afterwards if you see them in your TXT file.

Exporting results

When you are completely done correcting the recognition results, you can export the results to do e.g. a Forceed Alignment or to save the final texts. To start exporting, do the following:

1. Open the corrected XML file in WhisperCorector
2. Click on Settings; select 'Show Expert mode'
3. Click successively on 'Make Subtitles'; 'Export for FA'; 'Export Word Transcript'.

The results end up in the Results.

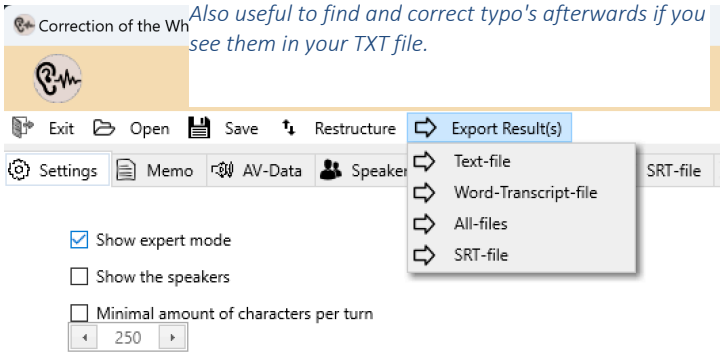

Fig. 11 Export of the corrected results

Explanation

- 'Make Subtitles' -> SRT file and can be used in video programmes
- 'Export for FA' overwrites the old TXT file with the corrected version and features time codes that are put in the right place by Forced Aligement (FA).
- 'Export Word Transcript', is found as _word.TXT file. You can open this and copy the contents to another word processing document. For example, in MS Word, you can select the pasted text and move the left margin to 3.5 or 4 cm, putting the speakers' names as entered in WhisperCorrector in front of the text.

Keep these .txt and _word.txt files, as they have the most accessible format for new analysis tools and are expected to remain readable in the future. In a copy, you can annotate the transcript. However, never change the original txt files so that they remain available for other applications. Keep all files together in the same folder so they can be used in programmes at the same time, such as a subtitle file and associated audio file.

Good practice

Some suggestions to ensure that transcriptions are as good and uniform as possible.

Sentences

Incomplete sentences end with three dots at the end.

For example:

"And then I thought..."

Those three dots indicate that the sentence ends, but that part of the text is actually missing.

Make sure the sentences are not too long, even if one long sentence is linguistically correct. For readability, it is preferable to split a long sentence into shorter ones.

For example:

"We are still working on it every day and I also want to convey that to people that most people who are going to watch this now will also agree with me that it is never finished."

Reads better as:

"We are still working on it every day. And that is also what I want to pass on to people, that most people who are going to watch it now will also agree with me that it is never finished."

Repetition

Whisper automatically omits unnecessary use of 'ehms'. Also avoid unnecessary repetition. You can limit frequent repetition to 3 times, for example, which reads more pleasantly and shows repetition just as well.

For example:

Instead of "Yes, yes, yes, yes, yes, yes, yes", note "Yes, yes, yes."

Punctuation

With proper and careful use of punctuation, spoken text can be made readable, so pay adequate attention to this.

Rules for punctuation differ from one language to another. Dutch has different rules from English. For English, see: The Punctuation Guide.

- Do not use loose punctuation marks, such as [space] - [space]. A floating punctuation mark will be interpreted as a word when aligning, causing text and audio to be out of sync.
- So: a comma is always 'attached' to the word, as here, as are 'inverted commas' and full stops. Including run-off dots, by the way...
- After ending dots, do not use the next punctuation mark, such as a comma or inverted comma. The sentence ended with 3 dots.
- Do not use leading dots.
- Do not use square brackets in the text [...], they are for the operation of the programme. You will find the speakers in the speaker display in the final transcript (word.txt), while the names are omitted in the subtitles (srt).