Oral History & Technology

Handleiding WhisperCorrector

Algemeen

WhisperCorrector-software (huidige versie: 1.4.2) is bedoeld voor het corrigeren van resultaten met Automatische Spraakherkenning (Automatic Speech Recognition - ASR) in srt en/of json formaat, zoals door Whisper. De software zet de herkende tekst (het bronbestand) om in een xml-bestand en toont de tekst naast het audio- of videobestand. Op dit moment kan Whisper nog geen sprekers herkennen of sprekerwisselingen aangeven. In WhisperCorrector kun je verschillende sprekers toe voegen, de tekst te corrigeren, één of meer woorden in één keer te vervangen, punten en komma's toevoegen, uitlooptekens achter onvoltooide zinnen, etc. Eenmaal klaar, kan de gecorrigeerde tekst in één keer of apart worden geëxporteerd naar 3 verschillende formaten:

OH-SMArt
ORAL HISTORY – STORIES AT THE MUSEUM AROUND ARTWORKS

Met ondersteuning van het PDI-SSH Project Oral History - Stories at the Museum Around Artworks (OH-SMArt)

1) een ondertitelingsformaat: SRT, kan onder video files geplaatst worden in video editing programma's
2) een simpel tekstformaat: TXT, kan gebruikt worden voor Forced Alignment*
3) een 'word.txt' formaat voor tekstverwerking in MS-Word, met opmaak zodat de linker kantlijn ingesteld kan worden op 3,5 cm inspringen, waardoor de sprekers voor de tekst komen te staan.

* Op dit moment, 24 juli 2023, is de FA-tool van het Centre for Language and Speech Technology van de Radboud Universiteit nog niet omgezet naar Whisper. WhisperCorrector v.1.4.2 heeft al een aardige mate van alignment geïntegreerd.

Whisper

Met het gebruik van de zogeheten End-to-End spraakherkenner zoals Whisper, is spraakherkenning behoorlijk goed geworden. De Nederlandse, Engelse, Duitse, Italiaanse en Spaanse geeft ±94% correcte herkenning van de gesproken woorden. Het loont dus om deze vorm van ASR te gebruiken en vervolgens eventueel de resultaten te corrigeren. Handmatig duurt transcriberen van 1 uur gesproken tekst, zo'n 6–8 uur. Met de combinatie van ASR en correctie daalt dit tot 1–2 uur, afhankelijk van de kwaliteit van de audio-opname, de duidelijkheid en juistheid van de gesproken taal, enz. Wat Whisper bovendien doet, is een redelijk correcte punctuatie. Punten, komma's, vraagtekens worden goed geplaatst en hoofdletters worden meestal gebruikt waar dat hoort.

Een klein nadeel van Whisper is dat herhalingen en niet afgemaakte zinnen niet worden weergegeven. Whisper, afkomstig van OpenAI, gebruikt waarschijnlijk het Large Language Model (GPT3 of GPT4) voor de herkenning wat ervoor zorgt dat je mooie zinnen terugkrijgt. Je kunt deze corrigeren met WhisperCorrector.

WhisperCorrector

WhisperCorrector is open source software, gratis te downloaden van https://speechandtech.eu als een zip-file (Windows) of een DMG-file (Apple). Het kan zijn dat toestemming gegeven moet worden op de computer, onder Systeeminstellingen/Privacy en beveiliging, om de gedownloade software te openen. Je kunt je materiaal op je eigen computer bewerken en opslaan in een veilige omgeving die je zelf kiest.

Start WhisperCorrector

Om te beginnen zul je éénmalig het transcriptieresultaat moeten importeren. Het bestandstype is afhankelijk van de versie van de Whisper-herkenner een srt- of json-bestand, beiden zijn goed. De eerste keer dat je WhisperCorrector opent, zie je een nogal leeg scherm (fig. 1).

Open éénmalig **Settings** om de juiste instellingen te selecteren om het herkende materiaal te importeren. De audio/video wordt meestal herkend indien deze in dezelfde map staat, en de files allen dezelfde naam hebben.

Oral History & Technology

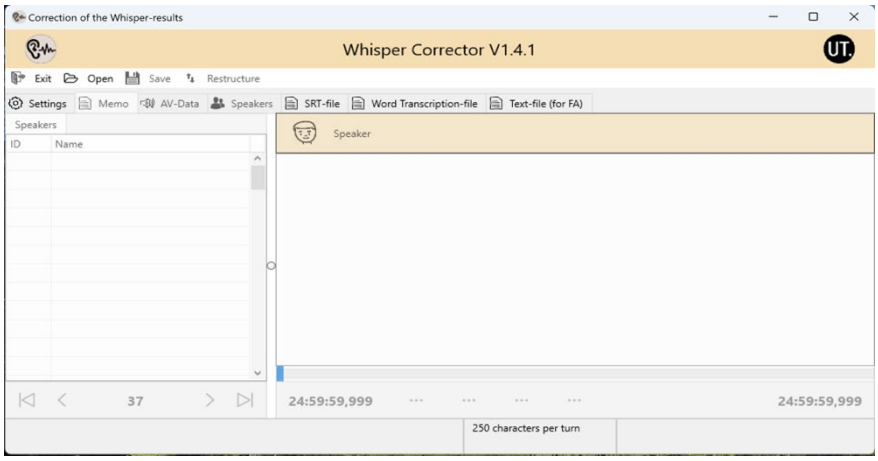

Fig. 1: WhisperCorrector versie (V 1.4.1) als je de software opstart.

Selecteer **Show expert mode** (fig. 2)

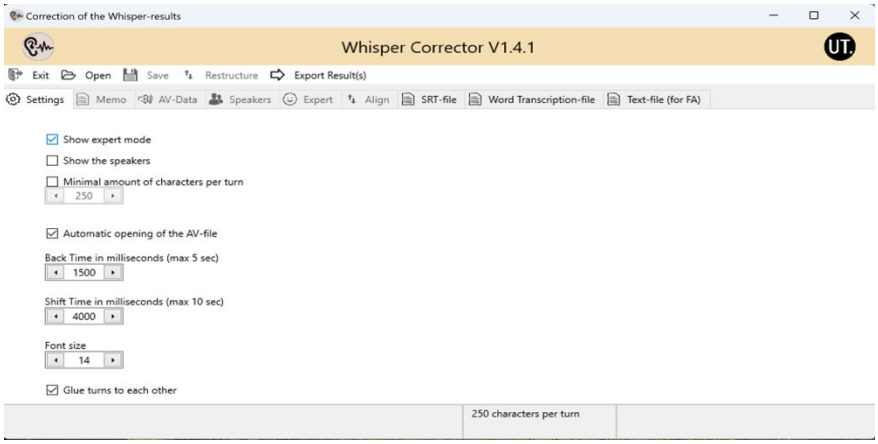

Fig. 2: Onder Settings kun je Show Expert Mode en Minimal Amount of Characters per Turn instellen.

Vul eventueel **Minimal amount of characters per turn** voor de hoeveelheid tekst in de memo-weergave, met minimaal 250 karakters of zoveel meer als wenselijk. Laat je het open, dan zal iedere tekstregel één 'turn' vullen. Als je een json-file inleest, dan worden standaard 250 karakters ingelezen (Fig.3). Kies voor langere stukken, bv 500 of 1000 karakters afhankelijk van je werkwijze. Combineren turns kan later ook nog wanneer je overgangen niet goed hoort bijvoorbeeld. Het totale aantal turns zie je onderaan op de mem-weergave(Fig.3), en kan flink oplopen wanneer korte stukken inleest.

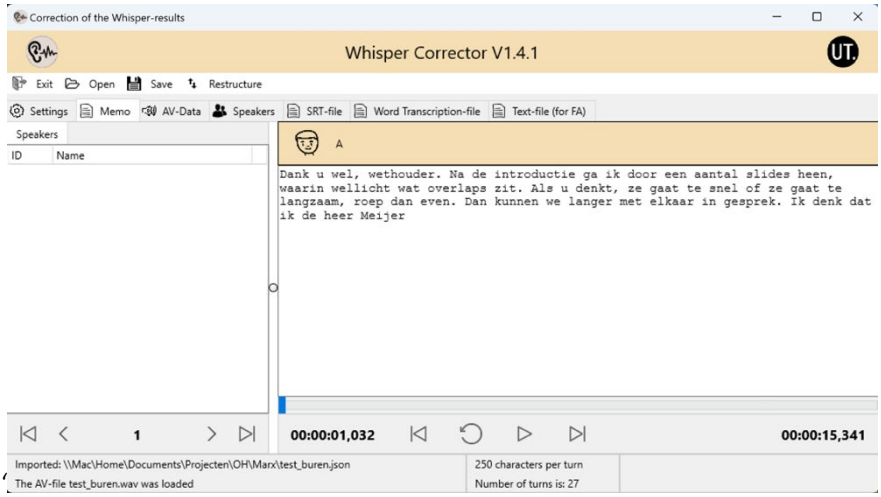

Fig. 3: importeren van een json- of srt-file. Default worden er bij een json-file ±250 karakters ingelezen

Oral History & Technology

Open JSON / SRT

Eenmaal in Expert mode, open de json- of srt-file. Het programma kijkt of er een bijbehorend sound of video-file beschikbaar is in dezelfde directory met dezelfde stamnaam. Als dat zo is, dan wordt die vanzelf geladen. Zo niet, open deze dan eventueel handmatig. Zorg dat deze dezelfde stamnaam heeft als alle andere files van de herkenner in dezelfde map staan. Klik 'Open', en selecteer **Soundfile** of **Videofile**, afhankelijk van hoe het is opgenomen.

Het kan zijn dat niet alle bestandsformaten gelezen worden, dan is omzetten naar wav altijd goed (ook video). Bij Windows moeten sommige gebruikers de video driver K-Lite Codec Pack Full installeren. Kies op de K-Lite Codec website 'server 1' (of 2 of 3) en installeer de software. Erna leest het programma de video-file meestal goed in.

Na het inlezen klik je op 'Save' en dan 'Save xml'. Het srt- of json-bestand heb je nu niet meer nodig. 'Show Expert Mode' is dan default weer uitgezet. (Je hebt deze stand nog één keer nodig wanneer je klaar bent, zie Resultaten exporteren).

Add Speakers

De laatste stap is sprekers toevoegen. Klik op de tab **Speakers** en vul voor iedere spreker de gegevens in (Fig. 4). Het veld **ID** wordt automatisch gegenereerd. De velden **Gender**, **Role** en **Description** hebben geen invloed op het correctieproces of de output en kunnen ook leeggelaten worden. Role: Interviewee of Interviewer. Het veld **Name** wordt overgenomen in het transcript zoals ingevuld. Het is handig deze niet onnodig lang te maken voor de leesbaarheid later, en het direct goed te doen. Verbeteren kan, maar indien je al ver met corrigeren bent en de spreker veel aan het woord is, moet het handmatig per turn worden bijgewerkt. De oude naam wordt niet meer als spreker herkend en komt met vierkante haken tussen de gesproken tekst terecht en moet daar dus verwijderd worden.

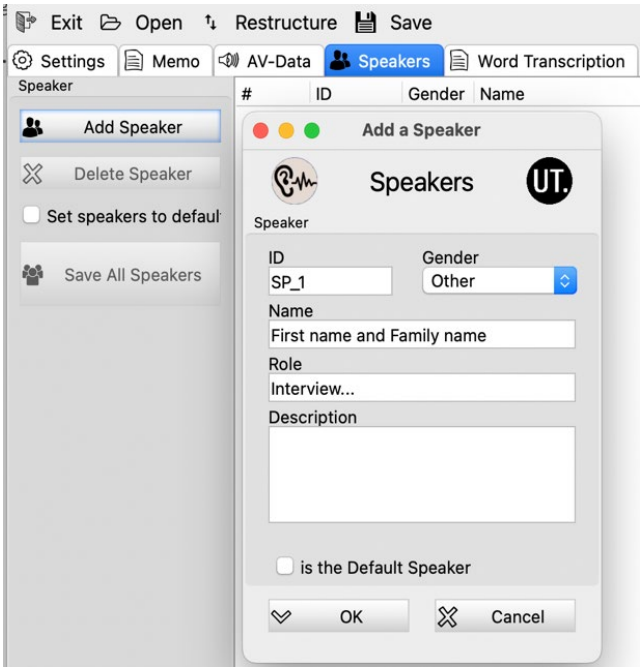

Fig. 4: Sprekers toevoegen.

De spreker die het meest aan het woord is kan als default gemarkeerd worden. Meestal is dat de geïnterviewde. Deze komt daarmee automatisch als spreker bovenaan iedere turn te staan (te wijzigen met rechtermuis klik). Alle sprekers ingevuld? Klik op **Save all Speakers**.

Save XML

Klik tenslotte helemaal bovenaan op **Save**, en kies **Save XML**. Vergeet dit nooit voor je afsluit, anders worden je bewerkingen niet opgeslagen.

Je hebt nu een xml-bestand gemaakt waarin je de automatisch herkende tekst kunt corrigeren. Het bronbestand hoeft nu niet meer geopend (in Expert mode), tenzij je helemaal opnieuw wil beginnen. Dan wordt een eerder gemaakte xml met eventuele correcties geheel vervangen en begin je weer van voren af aan.

Oral History & Technology

Open XML

Nadat je de srt/json-bestanden hebt ingelezen en samen met het audio/videobestand hebt opgeslagen als xml-bestand, begint de correctiefase van de spraakherkenning. Je opent vanaf nu steeds het xml-bestand – het werkbestand (Fig. 5). Je kunt altijd tussendoor opslaan en WhisperCorrector sluiten en later weer verder gaan. Klik **altijd** eerst op **Save -> Save XML** voordat je WhisperCorrector afsluit, om je bewerking te bewaren. Je bewerkte xml staat in de map op je computer, en niet in het programma. Je kunt aan meerdere transcripties tegelijk werken.

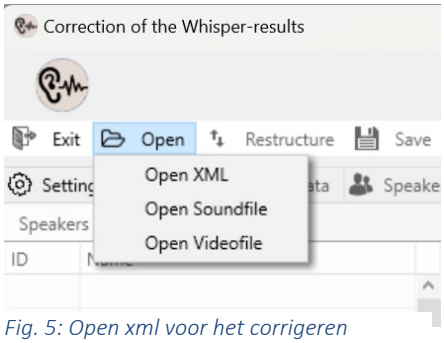

Fig. 5: Open xml voor het corrigeren

Correctie

Klik op Memo voor tekstverwerking van de ASR-resultaten. Rechts staat de herkende tekst waarin je kunt werken. Daaronder audioknoppen en de begin- en eindtijd van het fragment voor het afspelen/pauzeren van het geluidsfragment. De blauwe indicator in de balk er vlak boven kan naar een specifiek moment in het geluidsfragment worden geschoven. Bovenaan de tekst in de gele balk staat de default speaker, hier nog aangeduid met 'B', wat aangeeft dat de speakers niet zijn toegevoegd.

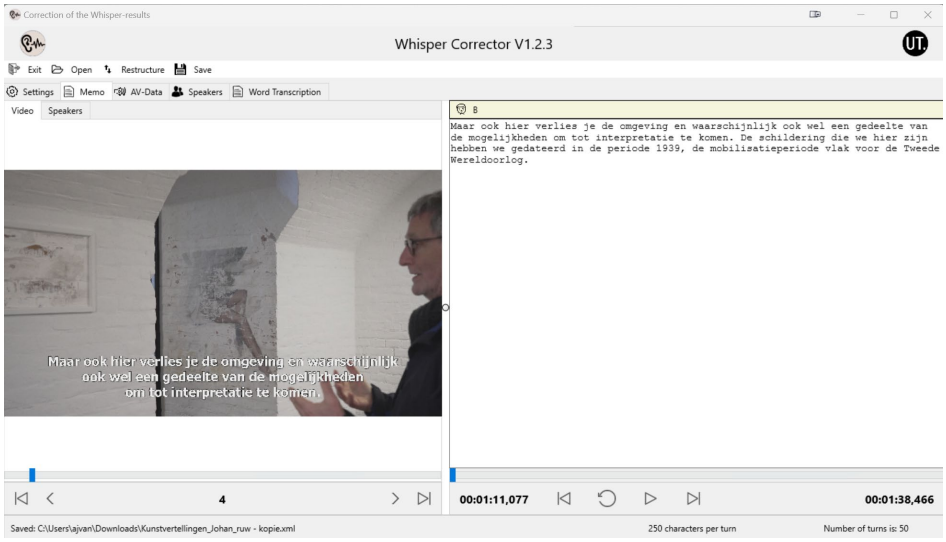

Fig. 6: Hier kun je horen en lezen wat er gezegd en herkend werd. In het tekstveld kun je corrigeren.

Links kun je uit 2 tabs kiezen, Video of Speakers. Daaronder knoppen om door het bestand heen te lopen, in grote stappen van -10 'turns' terug of +10 vooruit (buitenste tekens), of in stapjes van -1 of +1 (< of >) (Fig. 6).

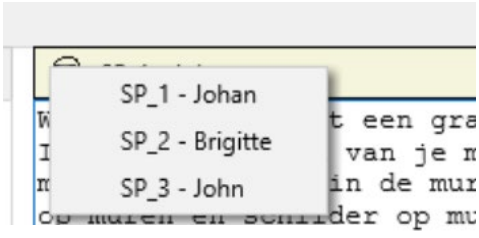

Fig. 7: Veranderen van hoofd spreker

Sprekers invoegen

De default spreker staat na invoeren bovenaan iedere turn. Kan veranderd door met rechter muisklik op het hoofd in de balk boven het tekstveld en selecteer de juiste spreker (Fig. 7). In het tekstveld zelf kun je ook een nieuwe spreker toevoegen. Ga naar de plaats in de tekst waar je een nieuwe spreker wil invoegen, rechtermuisklik, selecteer **Voeg een spreker in** en selecteer de juiste spreker. De software voegt de spreker toe met zijn ID en naam, bijvoorbeeld: [T3 - Peter Verboom] (Fig. 8). De harde returns hebben geen betekenis: ze zijn alleen toegevoegd om de leesbaarheid van de tekst te verbeteren.

Fig. 8: Voeg een andere spreker in het midden van de tekst

Combineren turns

Het kan gebeuren dat je wat je hoort wil combineren met de vorige of volgende turn. Dit gebeurt vooral wanneer de beurten vrij klein zijn. Klik daartoe met de rechtermuisknop op het nummer van de beurt in het rechter beneden venster en selecteer **Glue turn with prev one** of **Glue turn with next one** (Fig. 9). De 'gelijmde' tekst wordt samengevoegd in één tekstveld. Let op dat je eventuele correcties eerst opslaat, anders gaan die verloren in het lijmproces. Of eerst lijmen, en dan corrigeren.

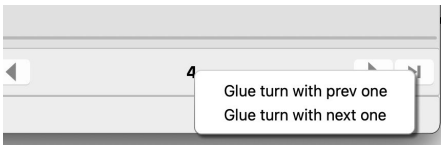

Fig. 9: Lijmen aan vorige of volgende turn

Zoek en vervang woorden

Een naam die niet met een hoofdletter is weergegeven of een veel voorkomende foutieve weergave kan in één keer per woord of woordcombinatie worden vervangen ter correctie. Selecteer daartoe het (de) te vervangen woord(en), klik met de rechtermuisknop op de selectie, en selecteer **Search & Replace** (Fig. 10). De software kopieert het woord in zowel het zoek- als het vervangveld. Als je op 'Ok' klikt, wordt het oorspronkelijke woord in het hele document vervangen door het nieuwe woord. Je kunt ook zonder het woord te selecteren zoeken en vervangen, bijvoorbeeld typo's die je bv naderhand in txt bestand ziet omdat ze rood onderstreept zijn. Dan hoef je ze niet op te zoeken.

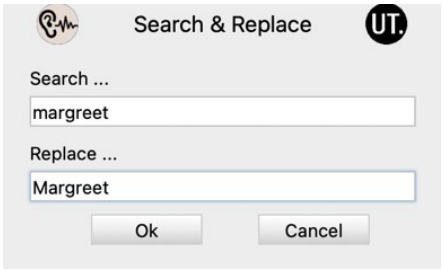

Fig. 10: Dialoogvenster van rechter muisknop functie voor Search & Replace. Ook handig om naderhand typo's corrigeren.

Resultaten exporteren

Wanneer je klaar bent met corrigeren van de herkenningsresultaten, kun je de verbeterde tekst exporteren en opslaan voor ondertiteling (SRT), Word-transcript en als "schone" tekstfile. Dat gaat als volgt:

1. Open gecorrigeerde xml, kies Settings
2. Selecteer 'Show Expert mode';
3. Klik op Export Results in format naar keuze, of All files in één keer (Fig. 11).

De resultaten belanden in de map met alle andere bestanden met dezelfde stamnaam.

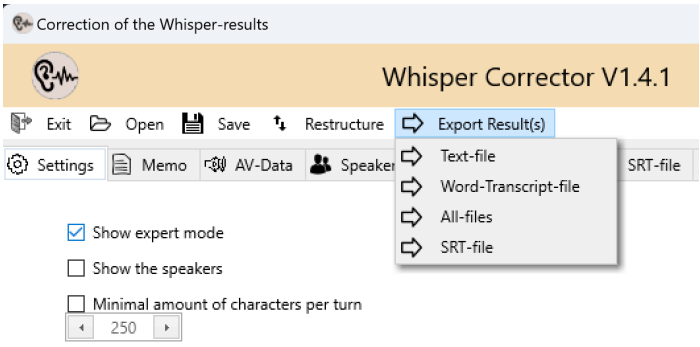

Fig. 11 Exporteren van de gecorrigeerde resultaten

Uitleg

- 'SRT-file', kan de oorspronkelijke srt-bestand vervangen met de gecorrigeerde versie en in videoprogramma's worden gebruikt. Voor een juiste overeenkomst tijd-tekst, kan de gecorrigeerde geschreven tekst gelijkgezet worden. Zie onder Algemeen, punt 2.
- 'Text-file' overschrijft het oorspronkelijke txt-bestand met de gecorrigeerde versie en is voorzien van tijdscodes die door Forced Alignment (FA) op de juiste plaats gezet moeten kunnen worden.
- 'Export Word-Transcript', vind je terug als _word.txt-bestand. Deze kun je openen, de inhoud kopiëren en plakken in een ander tekstverwerkingsdocument. In MS-Word kun je de geplakte tekst selecteren en de linker kantlijn voor inspringen verschuiven naar 3,5 of 4 cm, waardoor de namen van de sprekers zoals ingevoerd in WhisperCorrector voor de tekst komen te staan.

Bewaar de gecorrigeerde .txt- en _word.txt-bestanden ter archivering, zodat ze voor andere toepassingen beschikbaar blijven, want deze bestandsformaten blijven naar verwachting goed leesbaar voor andere taalanalyse tools. In een kopie kun je het transcript annoteren. Bewaar alle bestanden bij elkaar in dezelfde map, zodat ze in programma's tegelijkertijd in te lezen zijn.

Good practice

Enkele suggesties om ervoor te zorgen dat de transcripties zo goed en uniform mogelijk worden uitgevoerd.

Zinnen

Onvolledige zinnen eindigen met drie puntjes aan het eind.

Bijvoorbeeld:

> "En toen dacht ik..."

Met die drie puntjes geef je aan dat de zin eindigt, maar dat er eigenlijk een deel van de tekst ontbreekt.

Zorg ervoor dat de zinnen niet te lang zijn, ook al is één lange zin taalkundig correct. Voor de leesbaarheid, verdient het de voorkeur om een lange zin in kortere zinnen op te splitsen.

Bijvoorbeeld:

> "*We zijn er nog elke dag mee bezig en dat wil ik de mensen ook meegeven dat de meeste mensen die hier nu naar gaan kijken het ook met mij eens zullen zijn dat het nooit af is.*"

Leest beter als volgt:

> "*We zijn er nog elke dag mee bezig. En dat is ook wat ik de mensen wil meegeven, dat de meeste mensen die er nu naar gaan kijken, het ook met mij eens zullen zijn dat het nooit af is.*"

Herhaling

Whisper laat gebruik van 'ehms' automatisch weg. Vermijd ook onnodige herhaling. Veelvuldige herhaling kun je bijvoorbeeld beperken tot 3 maal, wat prettiger leest en waaruit de repetitie evengoed blijkt.

Bijvoorbeeld:

In plaats van "Ja, ja, ja, ja, ja, ja", noteer je: "Ja, ja, ja."

Interpunctie

Met goed en zorgvuldig gebruik van interpunctie is gesproken tekst goed leesbaar te maken, dus besteedt hier voldoende aandacht aan.

Regels voor interpunctie verschillen per taal. Het Nederlands kent andere regels dan het Engels. Voor Engels, zie: The Punctuation Guide.

- Gebruik geen losse interpunctietekens, zoals [spatie] **–** [spatie]. Een zwevend leesteken zal als woord worden opgevat bij het uitlijnen, waardoor tekst en audio niet meer synchroniseren.
- Dus: een komma zit altijd 'vast' aan het woord**,** zoals hier**,** evenals 'aanhalingstekens' en punten**.** Inclusief afloop puntjes, overigens**…**
- Na aflooppuntjes geen volgend leesteken gebruiken, zoals een komma of aanhalingsteken. De zin is met 3 puntjes beëindigd.
- Gebruik géén voorlooppuntjes.
- Gebruik geen vierkante haken in de tekst […], die zijn voor de werking van het programma. Je vindt de sprekers terug in de spreker weergave in het uiteindelijke transcript (word.txt), terwijl de namen juist weggelaten worden in de ondertiteling (srt).