SSHOC Webinar

CLARIN Hands-on Tutorial on Transcribing Interview Data

SSHOC
social sciences & humanities open cloud

SPEAKERS

Christoph Draxler

Researcher, Institute of Phonetics and Speech Processing at Ludwig Maximilian University Munich

draxler@phonetik.uni-muenchen.de

Henk van den Heuvel

Director Centre for Language and Speech Technology, Radboud University Nijmegen, Faculty of Arts

H.vandenHeuvel@let.ru.nl

NOTES

- **The webinar is being recorded.** All participants will receive a link to the recording later today.

- **Slides are available:** See the chat box for the link.

- **Questions?** Put them in the chat box. We'll put questions to the speakers at the end of the webinar.

SSHOC
social sciences & humanities open cloud

PREREQUISITES

- Eduroam account
  - Please make sure your eduroam account works – you will need it to access the OH Portal

- Audio sample files
  - `OH_Portal_demo.wav` test run of the webinar
  - `interview_Ben_Tucker.wav` interview with Ben Tucker
  - `Thompson-01.wav` interview with British general
  - `historic_interview.wav` analog socio-historical recording

- Google Chrome browser

Project:

SSHOC
social sciences & humanities open cloud

Horizon 2020
European Union Funding
for Research & Innovation

Type of action & funding:
Research and Innovation action
(INFRAEOSC-04-2018)

Partners: 45
(20 beneficiaries + 25 LTPs)

SSH ESFRI Landmarks and Projects
& international SSH data infrastructures

Project budget:
€ 14,455,594.08

Project website:
www.SSHopencloud.eu

Duration: 40 months
(January 2019 – 30 April 2022)

Objectives:

- creating the social sciences and humanities **(SSH)** part of European Open Science Cloud **(EOSC)**
- maximising **re-use** through **Open Science** and **FAIR** principles (standards, common catalogue, access control, semantic techniques, training)
- interconnecting existing and new infrastructures (clustered cloud infrastructure)
- establishing appropriate **governance model** for SSH-EOSC

EXPECTED IMPACT

The Social Sciences and Humanities are seamlessly integrated in the European Open Science Cloud

Availability of an EU-wide, easy-to-use SSH Open Marketplace, where tools and data are openly accessible

EU-wide availability of high quality "cloud ready" SSH tools and high quality SSH data

EU-wide availability of trusted and secure access mechanisms for SSH data, conforming to EU legal requirements

State of the art Research Infrastructure in several pilot domains advanced trough dedicated SSH data pilots cluster projects

Maximising reuse through Open Science and FAIR principles (standards, common catalogue, access control, semantic techniques, training)
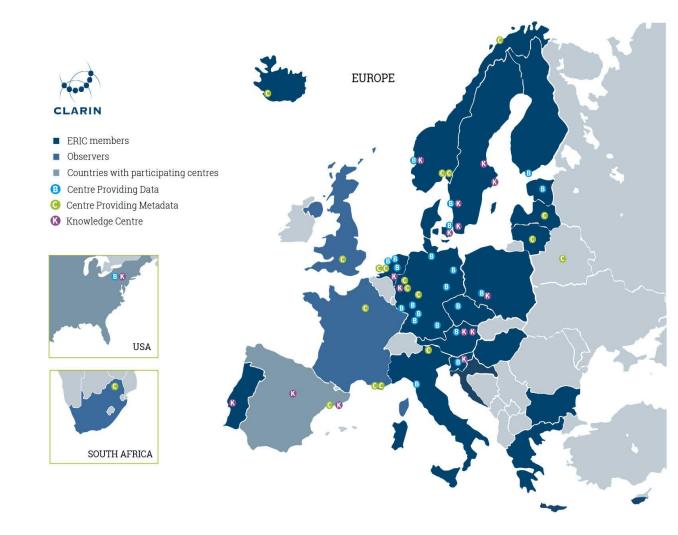
SSHOC
social sciences & humanities open cloud

CLARIN in six bullets

- **CLARIN** is the Common Language Resources and Technology Infrastructure
- **ESFRI** ERIC status since 2012, Landmark since 2016
- that provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form)
- and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
- through a **single sign-on** online environment

CLARIN ERIC in members and centres

A consortium of:

- 21 members: AT, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU, IS, IT, LT, LV, NL, NO, PL, PT, SE, SI
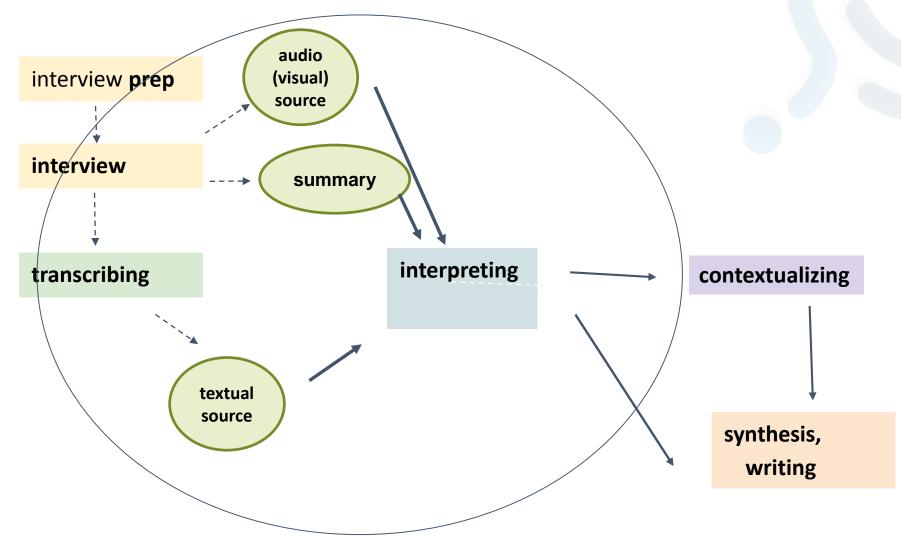- 3 observers: FR, UK, ZA
- >50 centres

CLARIN

- ■ ERIC members
- ■ Observers
- ■ Countries with participating centres
- Ⓑ Centre Providing Data
- Ⓒ Centre Providing Metadata
- Ⓚ Knowledge Centre

EUROPE

USA

SOUTH AFRICA

Technical pillars

- **Federated identity** - letting users login to protected data and services with their own institutional login and password

- **Persistent identifiers** - enabling sustainable citations of electronic resources

- **Sustainable repositories** - digital archives where language resources can be stored, accessed and shared

- **Flexible metadata and concept definitions** - to ensure semantic interoperability when describing language resources

- **Content search** - offering a search engine for a wide range of language resources

- **Web service chaining** - giving users the possibility to freely combine language processing services

CLARIN & ORAL HISTORY

- From Wikipedia, the free encyclopedia:
  - Oral history is the collection and study of historical information about individuals, families, important events, or everyday life using audiotapes, videotapes, or transcriptions of planned interviews. These interviews are conducted with people who participated in or observed past events and whose memories and perceptions of these are to be preserved as an aural record for future generations.

- Central are interview data

SSHOC
social sciences & humanities open cloud

What do oral historians do with interviews?

interview **prep**

interview

transcribing

audio (visual) source

summary

textual source

interpreting

contextualizing

synthesis, writing

SSHOC
social sciences & humanities open cloud

INTERVIEWS & LANGUAGE AND SPEECH TECHNOLOGY

🌀 Task force with CLARIN support:

https://oralhistory.eu/

CLARIN workshop in Arezzo in May 2017

Goals:
🌀 Draft a proposal for establishing an infrastructure directed at the orthographic transcription of existing OH-collections using Automatic Speech Recognition (ASR)
🌀 Show and practice various building blocks
   🌀 Before and during workshop
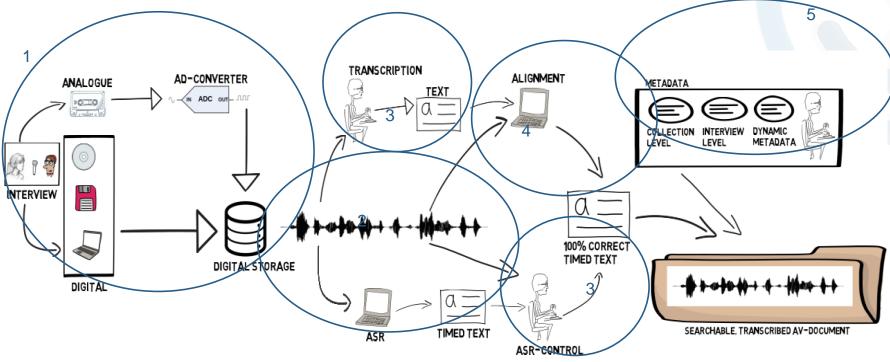🌀 Proposal for integration of building blocks into CLARIN infrastructure

Direct Result
🌀 Setting up the Transcription Portal at BAS Munich
🌀 Basis for a follow-up CLARIN workshops including (qualitative) analyses for which transcriptions are a precondition

SSHOC
social sciences & humanities open cloud

THE TRANSCRIPTION CHAIN

1. Analog-Digital Conversion (ADC)
   see https://oralhistory.eu/showcases/transcription-chain#digitizing-analogue-recordings
2. Automatic Speech Recognition (ASR)
3. Transcription correction
4. Synchronisation of Audio & transcription (Alignment)
5. Metadata

THE TARGET GROUPS

Starting in Arezzo, 2017:

- Historians and social science users who conduct research with recorded interview data sources;

- Linguists who use spoken language sources;

- Software tools specialists who develop support data processing and analysis tools.

Participants in next workshops (Munich, 2018; Utrecht, 2019) further included:

- Computational linguists

- Sociolinguists

- Social sign processing scholars

- Psychology, including psycholinguistics and mental health studies

THE TARGET GROUPS

But ..... think also of:

- Law scholars

- Social economists

- Empirical religion scholars

- Pharma-medical scientists

- Etc. Etc: everyone using interviews in their work

- What is your background?

Transcription Chain and the OH-Portal

- Transcription types
- Introduction to the T-Chain
- Demo: OH-Portal
- Recommended workflow

SSHOC
social sciences & humanities open cloud

Transcription Types

- Raw transcript

- Raw Transcript with diarization

- Transcript with some markup

- Broad transcript

- Man vs. Machine I & II

SSHOC
social sciences & humanities open cloud

Raw text transcript (Thompson-01)

my uhm argument for useful history is that actually all history starts by be
and tells someone about something and then they write it down yeah oka
the guy writing it down making it wrong or the guy may tell it wrong so jus
doesn't mean it is true uh I mean I I have commanded a brigade in a war
example war diaries are written now the British have a system called war
me from telling you something you know already no no no no every unit e
official diary of what they were doing every day okay and that is lodged in
now some academic historian look upon war diaries as being like you know the tablets of stone brought
down by by Moses it must be true because it is written down it is not I have come across war diaries where
there are actual lies being written because often a war diary would be written to uhm place the writers and
all these others in the best possible light yeah ok so it not necessarily accurate at all secondly things like
uhm after action reports uh pilots returning from missions the pilot is actually telling the intelligence officer
this is what I did okay it maybe backed up by photographs and things like that but it is still an oral account
patrols coming back will report to to the intelligence officer and it's an oral account so without wishing to
sort of bore you I actually think that oral history has a huge value provided you know what you are doing
when you are listening to it and you have a very good idea of who the people are and therefore how
trustworthy or other they are I think before I bore you to tears and and

This type of transcript is often used to evaluate the ASR recognition rate, or as input to other tools.

SSHOC
social sciences & humanities open cloud

Raw Transcript diarization

S1   my uhm argument for useful history is that actually all history starts by [...] someone about something and then they write it down

S2   yeah

S1   okay now

S2   back to to see this is

S1   so so the guy writing it down making it wrong or the guy may tell it wro[...] mean it is true uh I mean I I have commanded a brigade in a war the Falkla[...] are written now the British have a system called war diaries where every u[...] know already

S2   no no no no

S1   every unit every unit has to produce an official diary of what they were [...] in the national archive

S2   yeah

S1   now some academic historian look upon war diaries as being like you know the tablets of stone brought down by by Moses it must be true because it is written down it is not I have come across war diaries where there are actual lies being written because often a war diary would be written to uhm place the writers and all these others in the best possible light

S2   yeah

S1   okay so it not necessarily accurate at all secondly things like uhm after action reports uh pilots returning from missions the pilot is actually telling the intelligence officer this is what I did okay it maybe backed up by photographs and things like that but it is still an oral account patrols coming back will report to to the intelligence officer and it's an oral account so without wishing to sort of bore you I actually think that oral history has a huge value provided you know what you are doing when you are listening to it and you have a very good idea of who the people are and therefore how trustworthy or other they are I think before I bore you to tears and and

This type of transcript is often used
- **to perform automatic processing of the individual speakers' contribution,**
- **for content analysis, and**
- **as the basis for in-depth analyses**

Transcript with some markup

S1   my <F/> argument for useful history is that actually all history starts by being ~~~ someone about something and then they write it down

S2   yeah

S1   okay now

S2   back to to see this is

S1   so so the guy writing it down making it wrong or the guy may tell it wrong so ~~~ mean it is true <F/> I mean I I have commanded a brigade in a war the <NEGEO~~~ example war diaries are written now the British have a system called war diaries ~~~ from telling you something you know already</META>

S2   no no no no

S1   every unit every unit has to produce an official diary of what they were doing every day okay and that is <VOCAB>lodged</VOCAB> in now in the national archive

S2   yeah

S1   now some academic historian look upon war diaries as being like you know the tablets of stone brought down by by <NEBIB>Moses</NEBIB> it must be true because it is written down it is not I have come across war diaries where there are actual lies being written because often a war diary would be written <P 0.97/> to <P 0.46/> <F/> <P 0.96/> place the writers and all these others in the best possible light

S2   yeah

S1   okay so it not necessarily accurate at all secondly things like <F/> after action reports <F/> pilots returning from missions the pilot is actually telling the intelligence officer this is what I did okay it maybe backed up by photographs and things like that but it is still an oral account patrols coming back will report to to the intelligence officer and it's an oral account so <META>without wishing to sort of bore you</META> I actually think that oral history has a huge value provided you know what you are doing when you are listening to it and you have a very good idea of who the people are and therefore how trustworthy or other they are <META>I think before I bore you to tears and and</META>

This type of transcript is often used to mark regions of interest for further research, e. g. discourse analysis, emotion analysis

SSHOC
social sciences & humanities open cloud

Broad transcript

My argument for useful history is that actually all his[tory is] because somebody goes and tells someone about s[omething and you] write it down.

Just because it is written down it doesn't mean it is t[rue like war] diaries. The British have a system called war diaries [where you] produce an official diary of what they were doing eve[ry day. People and] historian look upon war diaries as being true becau[se they are written. They are] not! I have come across war diaries where there are [mistakes] because often a war diary would be written to place the writers in the best possible light. Secondly, things like after action reports. Pilots or patrols returning from missions tell the intelligence officer what they did, maybe backed up by photographs, but it is still an oral account.

I actually think that oral history has a huge value provided you know what you are doing when you are listening to it and you have a very good idea of who the people are and therefore how trustworthy they are.

This transcript is

- **close to the original recording, and**
- **it features punctuation, text smoothing etc. to make it easy to read.**

SSHOC
social sciences & humanities open cloud

Summary of Transcription Types

- The transcription types differ in terms of
  - verbatim fidelity
  - target audience
  - reusability
  - human expertise vs. machine processing needed

- In general, converting a
  - broad transcript to a more detailed one is not possible without the signal
  - very detailed transcript to a less detailed one may take as long as starting from scratch (unless tags can be removed automatically)

- Thus, a raw transcript with diarization is the most flexible solution

SSHOC
social sciences & humanities open cloud

Man vs. machine I (Thompson-01.wav)

my uhm argument for useful history is that actually all history starts by being oral because somebody goes and tells someone about something and then they write it down yeah okay now back to to see this is so so the guy writing it down making it wrong or the guy may tell it wrong so just because it is written down it doesn't mean it is true uh I mean I I have commanded a brigade in a war the Falklands war I know how for example war diaries are written now the British have a system called war diaries where every unit do stop me from telling you something you know already no no no no every unit every unit has to produce an official diary of what they were doing every day okay and that is lodged in now in the national archive yeah now some academic historian look upon war diaries as being like you know the tablets of stone brought down by by Moses it must be true because it is written down it is not I have come across war diaries where there are actual lies being written because often a war diary would be written to uhm place the writers and all these others in the best possible light yeah ok so it not necessarily accurate at all secondly things like uhm after action reports uh pilots returning from missions the pilot is actually telling the intelligence officer this is what I did okay it maybe backed up by photographs and things like that but it is still an oral account patrols coming back will report to to the intelligence officer and it's an oral account so without wishing to sort of bore you I actually think that oral history has a huge value provided you know what you are doing when you are listening to it and you have a very good idea of who the people are and therefore how trustworthy or other they are I think before I bore you to tears and and

my argument for the use of all history is a history Starts by being oral somebody goes until someone about something down in a war I know how for a God of War Diaries are written that the water unit stop every unit has produced in the diary what they were doing everyday ok and that lodged in our archive some tablets of stone brought down by by Moses it must be true I've come across Ward areas where they are our lives being written award will be written to place the writers in the best possible light after action report returning from missions this is what I did buy photographs and things and a cunt coming back no value provided you know what you're doing when you're listening and you have a very good idea who the people are and trustworthy to tell something

**Manual: 1779 chars
Google ASR: 717 chars
Levenshtein distance: 1075
~ 60.4 % error rate**

Manual transcript vs. Google ASR (default model, GB-English, called from BAS web services 17.02.2020)

SSHOC
social sciences & humanities open cloud

Man vs. machine II (OH_portal_demo.wav)

you now see some preliminary remarks they concern the application of the OH portal it's an application that runs in a browser window so as long as you don't close the browser you will always have access to the data that that you've been working on which is quite nice if something goes wrong you know the data is still there ok now we need to add some files so I will go to the operating system and select some files here is one by an English general and you can see now you can see the page of the OH portal there are four areas there is the top area where you can see a number of symbols they show what's going on at the moment they give you access to the help system they show you some statistics nice graphs showing how much time was spent doing ASR and how much time was spent doing manual transcription if you find anything strange in the software if anything doesn't work just click on the feedback button and send us a message the little bell tells you that something that you can switch on desktop notifications some processes may take a while so if you don't want to sit and stare at your screen but do something similar and do something more interesting than just click on the bell and it will the system will automatically notify if something's changed finally you have this button for settings then there's the second row with three blue buttons add files verify and start processing the middle area is reserved for messages so there you can see what's happening or what's going on at the moment in this case it says one audio file waiting to be verified then you see the column headings the column headers file upload speech recognition manual transcription word alignment per default speech recognition manual transcription and word alignment and phonetic detail are checked this means that our processing chain will go all the way from uploading the file through speech recognition you will do some manual transcription or correction and then finally there will be a word alignment and even phonetic details you might have seemed there is a little notification that came up and speech recognition actually has finished now the system or the file has arrived at the processing stage manual transcription and the symbol doesn't change to a little cog wheel instead has this little edit symbol let's go one step back first if you click on the icon on the on the green check mark next to the speech recognition or in the speech recognition column you can see that there are a number of options available already you can have the results of the speech recognition in one of the formats presented here so par is the BAS the Bavarian Archive for Speech Signals Partiturfile CTM is a format that's quite common for machine processing for speech recognition annot-json is new format that allows you to link the different elements on different annotation tiers which is very useful for later on processing TextGrid is a format might know when you've been working with Praat it shows the data in different tiers and on every tier all the items are marked with the precise time information table is a format that is very well suitable for R or Excel or any kind of statistics package and finally text is the raw text if we click on this we can see what Google made out of the audio file ok so this way you go through your file it might take a while until you've corrected everything in this case Google speech recognition results are not too good for a number of reasons probably the audio quality isn't too good although we can and understand it very well perhaps the topic that the speech recognition and engine is not familiar with we don't know that's one of the problems we access external services we have no idea what they actually offer and they only provide a very limited set of configuration options with which we could adapt the services to our needs ok I will close this and you can see the whole transcript let's imagine that the transcript was ok and we're done with the transcripts so all we do now is press on Send transcription meaning that now we have finished the manual transcription it is being sent to our server to the BAS server and now word alignment starts again this takes a few minutes and if you really have a very true manual transcript of what was said word alignment should give you very good results for the actual alignment of the words it will give you more it will give you alignment of the individual sounds but maybe this is something you don't need for your kind of research but I think the word alignment is something that you really need so let's have a look it's finished the word alignment already
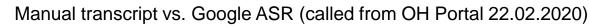
you know see some preliminary remarks they concern the application of the oh portal it's an application that runs in a browser window so as long as you don't close the door you will always have access to the data that you've been working on which is quite nice if something goes wrong you know the data is still there ok now we need to add some files so I will go to the operating system and select some files he has won by an English general and you can see now you can see the page of the oh portal they are four areas where is the top area where you can see a number of symbols they show what's going on at the moment they give you access to the help system they show you some statistics nice graphs showing how much time was spent doing as are and how much time was spent doing manual transcription if you find anything in the software if anything doesn't work just click on the feedback button and send us a message the little girl tells you that something that you can switch on desktop notifications some processes may take a while so if you don't want to sit and stare at the screen but do something similar and do something more interesting than just click on the bill and it will the system will automatically notify you the things changed finally you have the spartan phone settings then there's the second roll with three blue buttons and five verify and start processing the middle area is reserved for messages so there you can see what's happening or what's going on at the moment in this case it says one of your fire waiting to be verified then you'll see the column headings the column headers file upload speech recognition manual transcription word alignment the default speech recognition manual transcription and word alignment in phonetic detail are checked this means that our processing chain will go all the way from uploading through speech recognition you will do some manual transcription of correction and then finally there will be a word alignment and even phonetic details you might have seemed there is a little note that came up and speech recognition actually has finished now this is some of the file has arrived at the processing stage manual transcription and the symbol doesn't change to a little cog wheel instead has this little edit symbol let's go one step back first if you click on the icon on the on the green check Mark next to the speech recognition or in the speech recognition Collum you can see that there are a number of options available already you ca... far is the bus the Bavari... common for machine pr... different elements on di... performer might know w... all the items are marked... Excel or any kind of stat... Google made out of the... corrected everything Go... audio quality isn't too go... recognition and engine ... services we have no ide... options with which we c... transcript let's imagine that the transcript was ok and we're done with the transcripts are all we do now is press on Centrum scription meaning that now we have finished the manual transcription it is sent to our server to the bus service and I'll word alignment starts again this takes a few minutes and if you really have a very true manual transcript of what was said word alignment should give you very good results for the actual alignment of the word it will give you more it will give you the alignment of the individual sounds but maybe this is something you don't need for your kind of research but I think the word alignment is something that really need so let's have a look at finished the word alignment

Manual: 4641 chars
Google ASR: 4604 chars
Levenshtein distance: 248
~ 5.3 % error rate

Manual transcript vs. Google ASR (called from OH Portal 22.02.2020)

SSHOC
social sciences & humanities open cloud

ASR performance depends on audio quality and content

- Technical signal quality
  - sample rate
  - quantisation
  - channels
  - compression
- Recommendations
  - 44.1 or 48 kHz sample rate
  - PCM quantisation
  - 1 channel per microphone
  - .wav format

- Content
  - topic
  - speaking style & situation
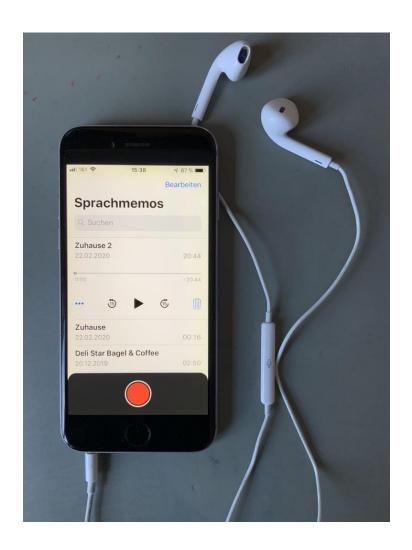 - environment noise
  - language, dialect, disorders
- Recommendations
  - aim for maximum speech & signal quality
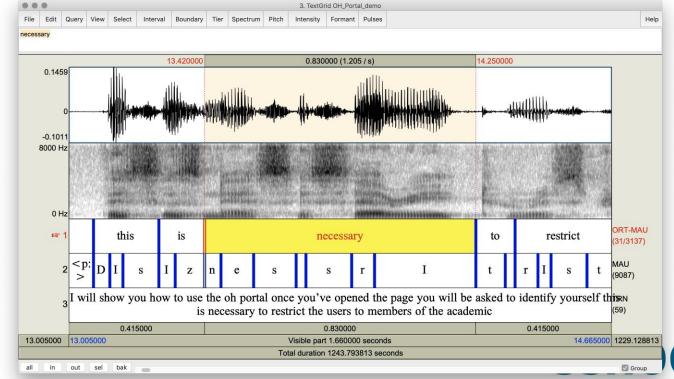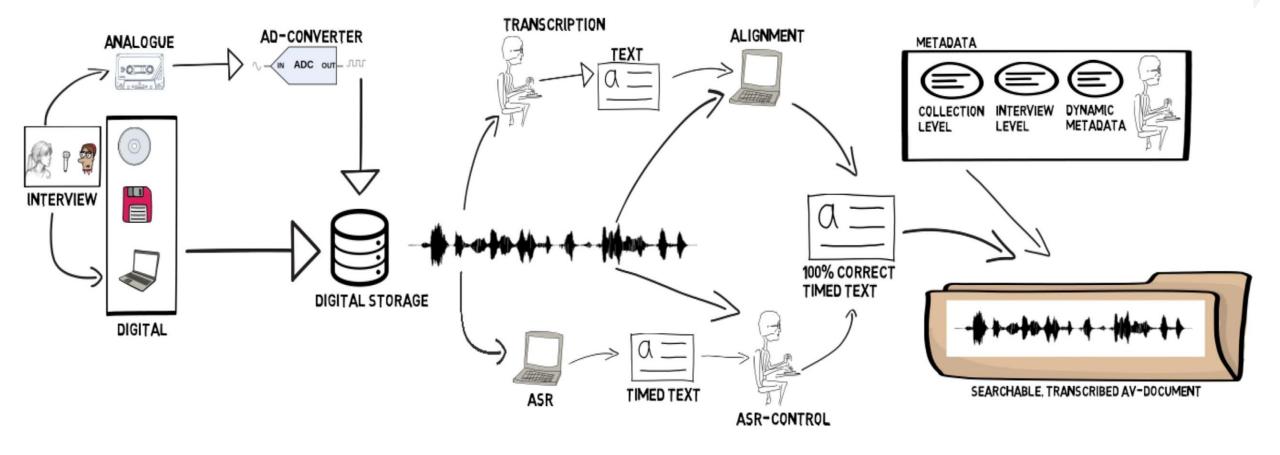 - close-talk or lapel microphones whenever possible

SSHOC
social sciences & humanities open cloud

Less may be more…

🌀 `OH_Portal_demo.wav` was recorded using an iPhone 6 with the standard headset microphone in a private living room

social sciences & humanities open cloud

T-Chain Overview

Automating the T-Chain

OH Portal

- Graphical interface to state-of-the-art speech processing tools
  - speech recognition
  - manual transcription
  - automatic word-alignment
- Preconfigured workflow
  - numerous export formats for further analysis
- Several languages
  - English, German, Dutch, Italian
  - others in the pipeline…
- Free for academic users
  - requires authentication

OH Portal: Recommendations

- Currently, the OH-Portal has some limitations
  - maximum file size approx. 200 MB
  - .wav audio file format
- Thus, we recommend to
  - split your audio file into meaningful units (5-10 mins), e.g. introduction, housing situation, life at home, work, leisure time, etc.
  - use e.g. Audacity to
    - split stereo files to two mono files
    - downsample to 16 kHz
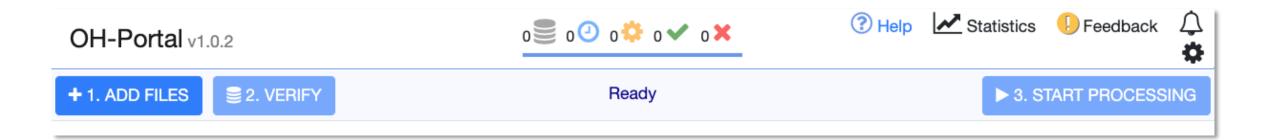   - convert format to .wav
- Don't use special characters in your file names!

SSHOC
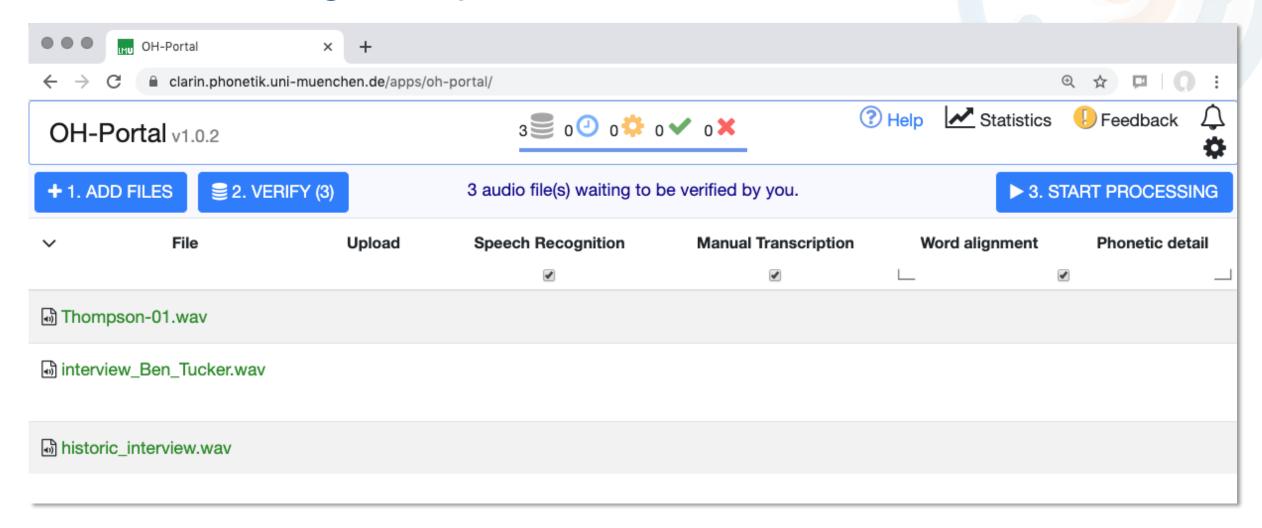social sciences & humanities open cloud

OH-Portal v1.0.2

0 ≡ 0 ⏱ 0 ⚙ 0 ✔ 0 ✖

? Help | 📈 Statistics | ❗ Feedback | 🔔 | ⚙

+ 1. ADD FILES | **≡ 2. VERIFY** | Ready | **▶ 3. START PROCESSING**

⌄	File	Upload	Speech Recognition	Manual Transcription	Word alignment	Phonetic detail
			☑	☑		☑	

Quickstart

1.

Add files

You can drag & drop files (.wav & transcripts) to this whole area (inside the dashed rectangle) or you can use the "1. ADD FILES" button. Later, you can add files via drag & drop to the table rows, too.

2.

Verify new files

Before the new files can be processed it's required to set few options. Click on the "2. Verify" Button.

3.

Start Processing

After all files are verified you can click on "START PROCESSING". The application starts the processing of pending tasks.

SSHOC

social sciences & humanities open cloud

Using the OH Portal

- Top row: version number, status icons, etc.
  - please use the feedback button to report problems
- Action buttons
  - start and stop processing

OH Portal: drag & drop audio files
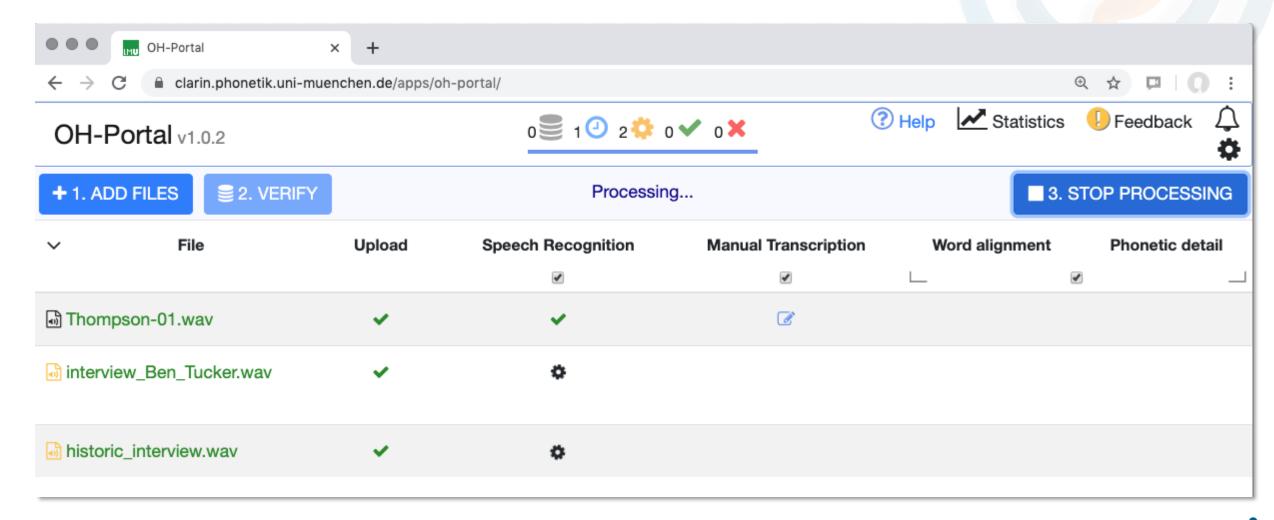
OH Portal: select ASR service

Click on logo to view terms of use

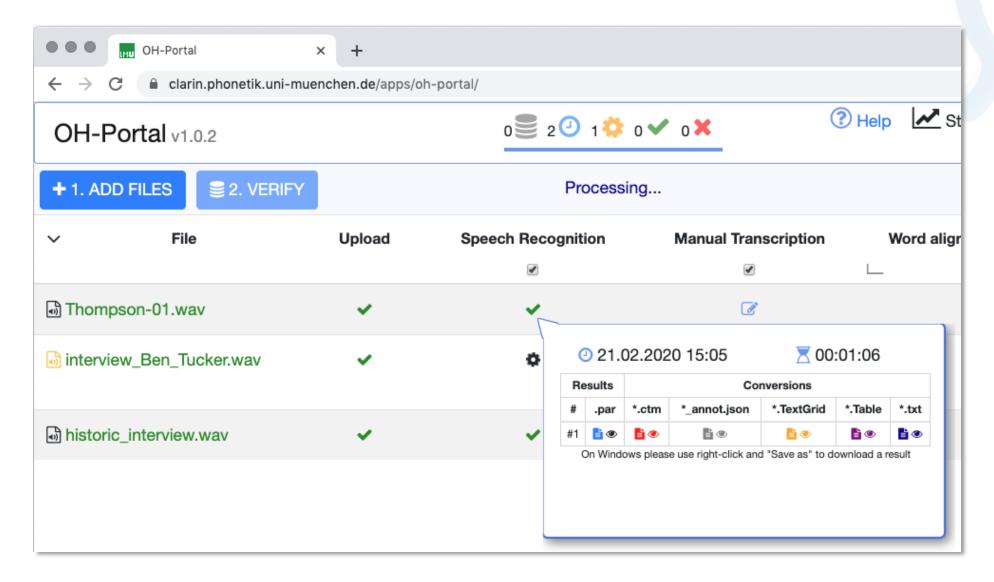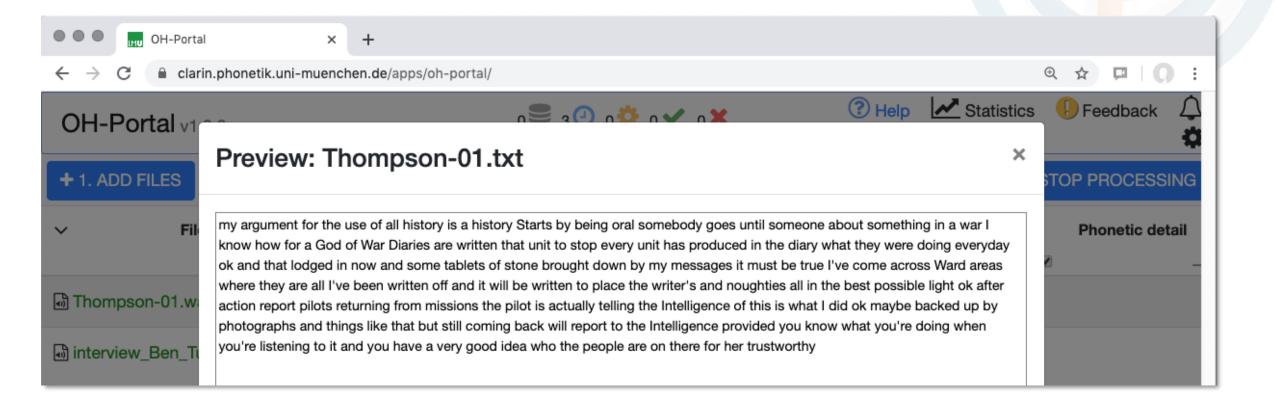☐ service providers may update services and terms of use without prior notice!

SSHOC
social sciences & humanities open cloud

OH Portal: monitor progress

OH Portal: download options

OH Portal: Preview result data

OH Portal: manual correction

- built-in transcription editor Octra
  - press tab to start and stop audio
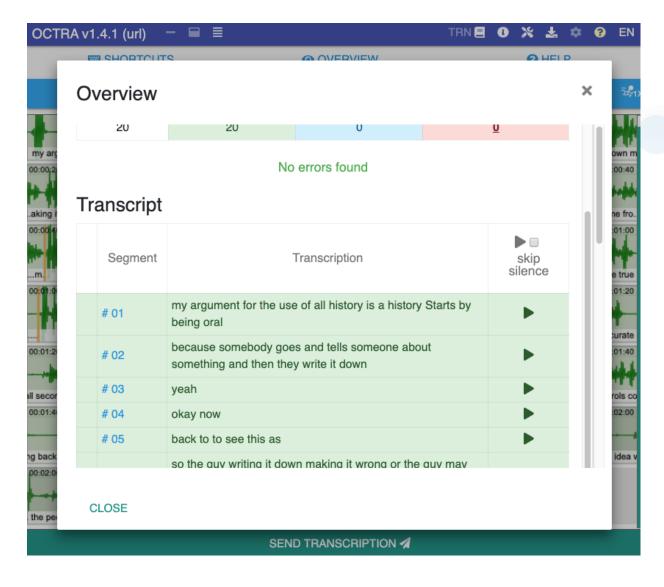 - press return to open segment for transcription

OH Portal: transcription units

OH Portal: quick check with overview editor

OH Portal: finally, download transcripts

Download results by column ✕

This creates a zip-archive of all the selected results and optionally conversions to other formats. If you do not select any additional conversions only the original results are added to the zip-archive.

Add conversions (optional):

☐ 📄 CTM (.ctm) ☐ 📄 BAS Partitur Format (.par)

☐ 📄 AnnotJSON (_annot.json) ☐ 📄 TextGrid (.TextGrid)

☐ 📄 Table (.Table) ☐ 📄 Plain Text (.txt)

Get package

Close

- File-wise, by clicking on the check mark
- Batch-wise, by clicking on the column head

SSHOC
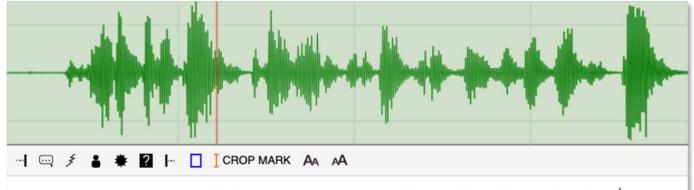social sciences & humanities open cloud

Recommended workflow

- Split recording into meaningful units
- Run ASR on recordings
- Check ASR results
  - 👍: skip manual correction
  - 👎: do manual correction
- Run word-alignment
- Download transcript(s)

SSHOC
social sciences & humanities open cloud

Where to now?

because somebody goes and tells someone about something and then they write it down

SSHOC
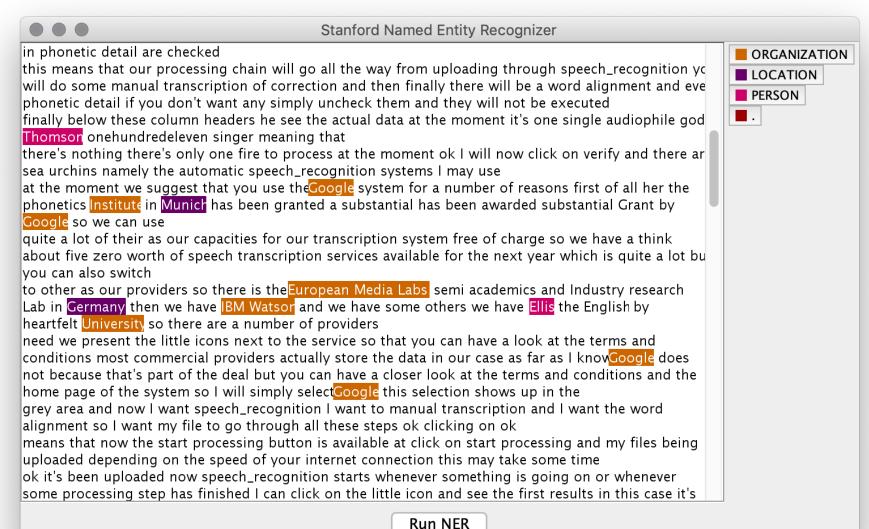social sciences & humanities open cloud

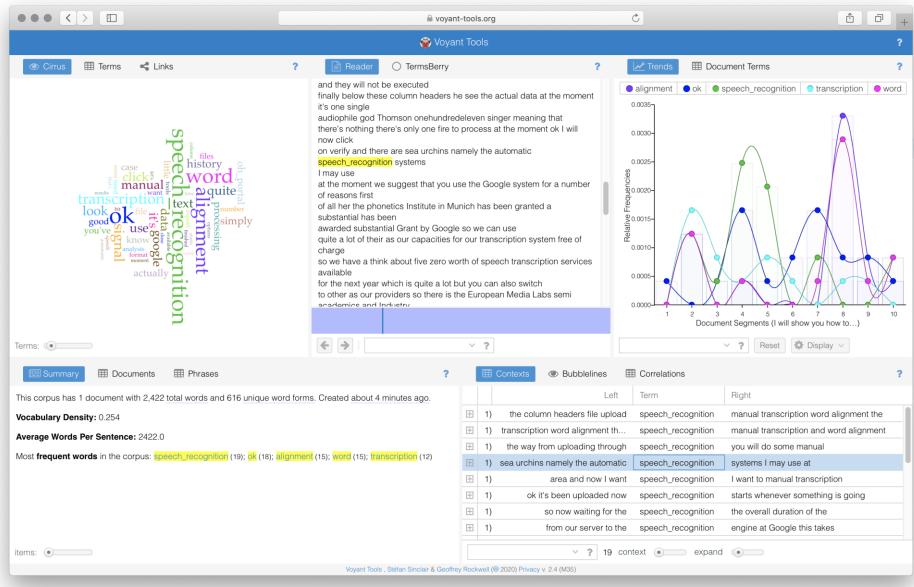# In-depth text analysis: Named Entity Recognition

Screenshot of Stanford Named Entity Recognizer (v. 3.9.2)

Used with the included basic 3 class NER tagger trained on standard text corpora

In-depth text analysis (e. g. voyant-tools.org)

UPCOMING EVENTS AND WORKSHOPS

- SSHOC Consortium Meeting, March 24, 2020, Utrecht: **SSHOC Workshop - Linking Social Survey and Linguistic Infrastructures through EOSC**
    - https://www.sshopencloud.eu/sshoc-workshop-linking-social-survey-linguistic-infrastructures-eosc
    - First SSHOC ASR Use Case on panel interview data, pilot
    - (CLARIN) tools for analysis of audio

- **LR4SSHOC** Workshop @LREC, May 11, 2020, Marseille:
    - https://www.clarin.eu/LR4SSHOC
    - Cross disciplinary tools for analysis of interview data

- ICMI Workshop, 25-29 October 2020, Utrecht: **Speech, Voice, Text, and Meaning**
    - http://icmi.acm.org/2020/
    - A Multidisciplinary Approach to Interview Data through the Use of Digital Tools

SSHOC
social sciences & humanities open cloud

Thank you for your attention!

Questions?

Please put them in the chat box.

Slides and a recording will be sent to all registered delegates.

Join our community

https://www.sshopencloud.eu

info@sshopencloud.eu

@SSHOpenCloud

/in/sshopencloud