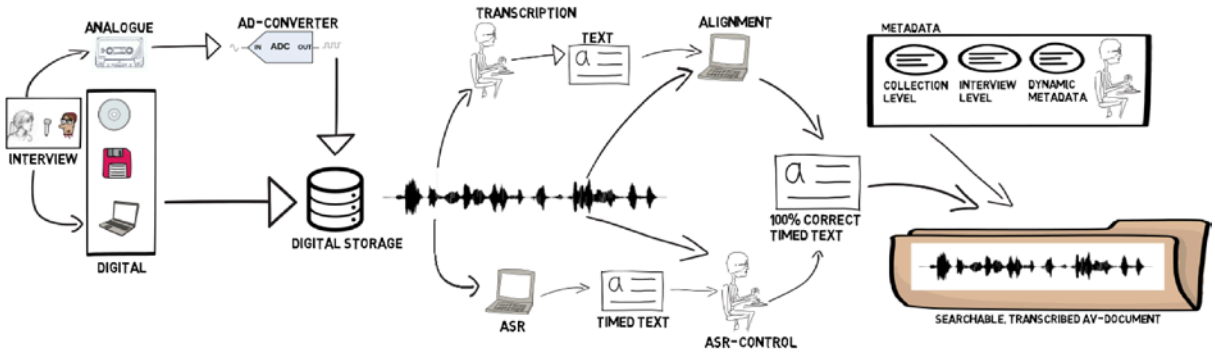
Report CLARIN Oral History Workshop Arezzo



A REPORT ABOUT THE OH-WORKSHOP IN AREZZO, THE COMMENTS OF THE VARIOUS PARTICIPANTS AND THINGS TO DO FOR THE PROPOSAL

HENK VAN DEN HEUVEL

ARJAN VAN HESSEN

Report CLARIN Oral History Workshop Arezzo

ORGANISING TEAM

HENK VAN DEN HEUVEL	CLST RADBOUD UNIVERSITY
ARJAN VAN HESSEN	Utrecht University / University of Twente
SILVIA CALAMAI	DSFUCI, University of Siena
LOUISE CORTI	UK Data Archive, University of Essex
STEF SCAGLIOLA	University of Luxembourg
MARTIN WYNNE	IT Services, University of Oxford

LOCATION AND DATE

Arezzo, Italy, 10-12 may 2017

The location for the workshop was the Department of Education, human sciences and intercultural communication – Siena University, Campus 'Il Pionta' (*UNISI-Dpt. Scienze della formazione, umane e della comunicazione interculturale*). See http://oralhistory.eu/workshops/arezzo#location.

All material related to the workshop, including presentations given, are collected in a folder in Google Drive.

INTRODUCTION AND GOAL

This type II workshop proposal is a follow-up of a type I CLARIN workshop organised in Oxford[1] on 18-19 April 2016.

Among other things this workshop demonstrated the widely-supported demand for effective and cost-efficient ways to make transcriptions of OH interviews. An action line from the minutes of the workshop is directed towards establishing a production line from audio to orthographic transcriptions of sufficient quality for OH researchers to proceed to the next text-based processes in their analysis paradigms.

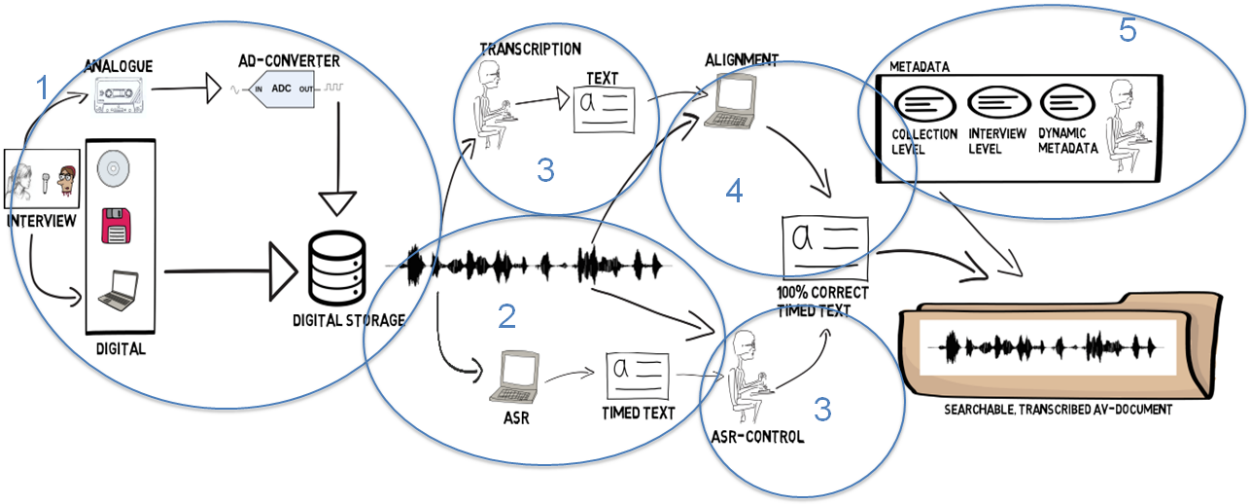
In this pilot, we aim to bring the components together for three languages: Dutch, English and Italian.

This transcription chain consists of five components outlined below:

1. **Analogue-to-digital (AD):**
   Digitising analogue material in such a way that it resembles the original audio quality AND becomes optimally suitable for ASR.

[1] https://www.clarin.eu/event/2016/clarin-plus-workshop-exploring-spoken-word-data-oral-history-archives

2. Automatic Speech-to-Text (ASR):
   Making a service where non-technical people easily can upload AV-recordings and retrieve the recognition results.
3. Transcription:
   Offering a web-based service where scholars can enter transcriptions and/or adjust the errors made by the ASR-engine e.g. in a crowdsourcing environment
   3A. Transcription guidelines
4. Alignment of speech and text:
   Offering webtools that take audio and transcriptions as input and synchronise these for easy playback and transcription correction
5. Metadata:
   Provide interface to add metadata about the recording. When the metadata files adhere to standards then the interviews they refer to become searchable and can be processed with digital tools.



The workshop should result in a proposal for establishing an infrastructure directed at the orthographic transcription of existing OH-collections, through the integration of existing modules. This aim has two sub-goals: A. to achieve a guideline document to be shared on the web with an implementation of this transcription chain. B. to develop an essential part of the transcription chain such that it has a true added value for the researchers and can be implemented in 3 PM.

A preparatory workshop with the coordinating team was organised in Utrecht on December 6 and 7, 2016. In this workshop, which was held at the SURF premises, we detailed the envisaged transcription chain, made an agenda for the workshop in Arezzo and fixed a date for this, and we selected targeted participants.

PARTICIPANTS

An overview of the 21 participants from 5 different countries. More detailed information can be found here.

Country	Expertise	Name	email
IT	OH	Silvia Calamai	silvia.calamai@unisi.it
IT	OH	Bianca Pastori	pastori.bianca@gmail.com
IT	TECH	Piero Cosi	piero.cosi@cnr.it
IT	INFRA	Riccardo Del Gratta	riccardo.delgratta@ilc.cnr.it
IT	INFRA	Monica Monachini	Monica.Monachini@ilc.cnr.it
UK	OH	Louise Corti	corti@essex.ac.uk
UK	OH	Graham Gibbs	g.r.gibbs@hud.ac.uk
UK	OH	Maureen Haaker	mahaak@essex.ac.uk
UK	TECH	John Coleman	john.coleman@phon.ox.ac.uk
UK	TECH	Thomas Hain	t.hain@sheffield.ac.uk
UK	INFRA	Martin Wynne	martin.wynne@bodleian.ox.ac.uk
NL	OH	Stef Scagliola	stefania.scagliola@uni.lu
NL	OH	Norah Karrouche	karrouche@eshcc.eur.nl
NL	OH	Afelonne Doek	ado@iisg.nl
NL	TECH	Arjan van Hessen	A.J.vanHessen@utwente.nl
NL	TECH	Henk van den Heuvel	h.vandenheuvel@let.ru.nl
NL	INFRA	René van Horik	rene.van.horik@dans.knaw.nl
NL	TECH/INFRA	Roeland Ordelman	rordelman@beeldengeluid.nl
DE	TECH/INFRA	Christoph Draxler	draxler@phonetik.uni-muenchen.de
CZ	INFRA	Pavel Stranak (LINDAT)	stranak@ufal.mff.cuni.cz
NL	VIDEO	Leon Wessels	l.c.wessels@uu.nl

AGENDA

WEDNESDAY 10 MAY:

14:00	Welcome	Silvia Calamai	
14:15	Overview	Henk van den Heuvel	Background, Objectives, Agenda, targets of workshop
14:30	Transcription chain	Henk van den Heuvel	The various building blocks of a transcription chain, as discussed in Utrecht workshop.
14:45	AD-conversion	Arjan van Hessen	AD-conversion-tools
	ASR-TOOLS: FULL SPEECH RECOGNITION FOR DIFFERENT LANGUAGES		
15:00	ASR tools, English	Thomas Hain	Focussing at WebASR.org
15:20	ASR tools, Dutch	Roeland Ordelman	KALDI recognizer Dutch NISV
15:40	ASR tools, Dutch	Henk van den Heuvel	Webinterface incl. OH version, incl results
16:00	BREAK		
	ASR-TOOLS: ALIGNMENT OF AUDIO AND TRANSCRIPTS FOR VARIOUS LANGUAGES		
16:15	WebMAUS	John Coleman & Christoph Draxler	WebMAUS Aligner
16:30	Italian Alignment	Piero Cosi	The Italian Aligner
16:45	Experience feedback	Graham Gibbs	Participants reports on their experiences with the ASR tools and Alignment tools
17:15	DIY	Arjan van Hessen	Discussion about desired formats of the ASR-tools. What do you want to get back from the ASR-engine? Hands-on Experience if necessary
18:30	Close of day 1	Silvia Calamai	Are you hungry?
19:30	Dinner		

THURSDAY 11 MAY:

9:15	Buon Giorno	Henk van den Heuvel	Summary of day 1 and Overview of day 2
	Transcription: Guidelines, Standards, Editors, Crowdsourcing		
9:25	Transcription guidelines	Stef Scagliola & Silvia Calamai	Various standards, best practices for Oral History
9:45	Manual transcription correction services	Arjan van Hessen	What is there to be used by individual researchers (for example SubtitleEdit)
10:00	Web-based annotation editors	Christoph Draxler	Portal for individual researchers and in in a crowdsourcing environment
11:00	BREAK		
11:15	Crowdsourcing	Arjan van Hessen	Crowdflower crowdsourcing strategies and transcription correction
11:25	Discussion	All	Participants reports on their experiences with Transcription services and crowdsourcing platforms
12:00	Hand-on experience	Arjan van Hessen & Christoph Draxler	Do a correction of your own transcriptions, set-up a crowdsourcing experiment where people can help you with the transcriptions, and try-out the transcription guidelines (good or not and what is missing)
13:00	LUNCH		
	METADATA: GUIDELINES, STANDARDS, EDITORS		
14:00	Metadata	Stef Scagliola & Louise Corti	Overview of standards, relevant categories, language of metadata, translation etc
14:30	Metadata editor	Henk van den Heuvel	A metadata editor as implemented at CLST
14:45	Discussion	All	Participants reports on their experiences with Metadata-editing
15:00	BREAK		
	PRESENTATIONS ON DATA MANAGEMENT/HOSTING IN NL, UK, IT ((PERSISTENT) ARCHIVING OPTIONS)		
15:15	National Infra: **NL**	Rene van Horik	About the data infrastructure in the country and how our services could fit into that & access to data, tools, metadata for the research community at large & IPR / informed consent / ethical issues

15:30	National Infra: UK	Louise Corti	About the data infrastructure in the country and how our services could fit into that & access to data, tools, metadata for the research community at large & IPR / informed consent / ethical issues
15:45	National Infra: IT	Monica Monachini	About the data infrastructure in the country and how our services could fit into that & access to data, tools, metadata for the research community at large & IPR / informed consent / ethical issues
16:00	National Infra: CZ	Pavel Stranak	About the data infrastructure in the country and how our services could fit into that & access to data, tools, metadata for the research community at large & IPR / informed consent / ethical issues
16:20	Discussion	Henk van den Heuvel	
18:00	Close of day 2	Silvia Calamai	

FRIDAY 12 MAY:

9:45	Buongiorno	Henk van den Heuvel	Summary of day 2 and overview of day 3
10:00	Wrapping up	Henk van den Heuvel	● Which improvements are needed for the Google documents on the various topics: ● Which software improvements are needed and should be included in the implementation plan ● Which facilities do we miss so far?
10:30	Proposal	Arjan van Hessen	Concluding actions for finalising the implementation proposal
11:30	BREAK		
11:45	Time schedule	Arjan van Hessen	Setup of the time schedules for the next months: from workshop to proposal.
12:15	Plan for a publication	Stef Scagliola	How to set up some publications based on the work done in this workshop?
12:45	LUNCH		
14:00	Adjourn	Henk van den Heuvel & Silvia Calamai	

Presentations

All presentations can be found in this folder.

Homework

In the week before the workshop we asked our participants to try out several tools and to report about their experiences during the workshop. The instructions for the homework can be found here.

Participants could report their experiences with the tools in a Google spreadsheet. Not everyone reported about all tools, but together, the responses gave a good impression about their experiences.

Notable issues from the discussions

There are detailed notes from the presentations and discussion at the workshop. Here follows an election of remarks considered relevant:

The Transcription Chain

The Transcription Chain is not linear. For an optimal result re-iterations are necessary. Also, combinations are relevant: Evaluation by comparison of output from various models can be helpful. E.g. Comparing output from aligner and ASR helps to decide for which interviews (or parts) the tools can speed up the transcription process.

Moreover, the transcription chain needs various format converters for:

audio (audio formats used by the scholars -> audio formats accepted by the tools) and for the transcriptions: all tools produced a different output.

ASR

Specialised OH-versions for ASR are relevant to recognise domain specific content words and names. Moreover, there needs to be an option for the users to add their own vocabularies. One may think about list of special names, places and organisations, mentioned in the interviews

There is also a wish to add pause durations to the output transcriptions.

Much OH-speech is vernacular and/or contains dialect. To recognise these non-standard interviews, special recognisers, trained with dialect speech, should be build by the national ASR-providers.

Hidden facilitie

There are many hidden facilities in the various tools:

- WebASR for UK English:
  aux data improves recognition; more output options: speaker turns
- OH-ASR for Dutch:
  options for XML-output and noise symbols in transcriptions can easily be added.

Manuals

An overall complain (from the humanity scholars) was the lack of good, clearly written manuals, explaining the "*dos and don'ts*" of the tools. To much is background knowledge is expected by the computer scientists that made the systems.

- Infrastructure oriented approach is needed

-Registration of collection/recordings by owners

-Standardisation of formats including metadata

TRANSCRIPTIONS

Recommendations for transcriptions:

- Introduce time codes at utterance level as basis
- Codes for
  - Speaker turns
  - Nonverbal, pauses silences
  - Interruptions, false starts
  - Dialects indication
  - Comments
  - Customised elements

Transcription tools:

an overview of the various features of recommended transcription tools should be added to our guidelines of the transcription chain. => Homework for speech technologists

Respeaking:

Respeaking in the same and/or another language can greatly improve ASR output yielding multiple audio and transcription streams

OCTRA:

The OCTRA tool is a web-based transcription tool that can be used by one or more people (a kind of unpaid crowdsourcing platform)

Take into account http://oralhistorianstoolbox.cohds.ca/

METADATA

Metadata-editor at http://applejack.science.ru.nl/oh-metadataregistry :

- At collection level or at item level. Both should be distinguished
- PID at collection level
- Linking to other resources: audio etc.
- Add description with examples for each of categories
- Also, look at COMEDI http://clarino.uib.no/comedi/page

Consider storing all data related to a recording together like in AESS Data Entry

Language used in attributes. Must it be English? Preferably.

INFRASTRUCTURE FOR "T-CHAIN"

In the current situation with the limited amount of resources for the 3 PM developer's time we should go for a federated approach, rather than integrated approach where everything is under one roof in one ICT-environment. Once an integrated approach becomes possible, we should step in there. But not our target now.

As a group, we should try to meet again after 3 years and review the state of affairs.

FOLLOW-UP PROPOSAL

The most crucial block in the chain where technology can help the researchers is the ASR component. A versatile interface hosting ASR engines for several languages is considered the key focus point for our efforts given the fact that we have only 3 PM developers time to spend.

At the last day of the workshop Arjan van Hessen presented a basic idea of how to arrange this.

The current situation of dispersed and partly unavailable ASR recognisers:

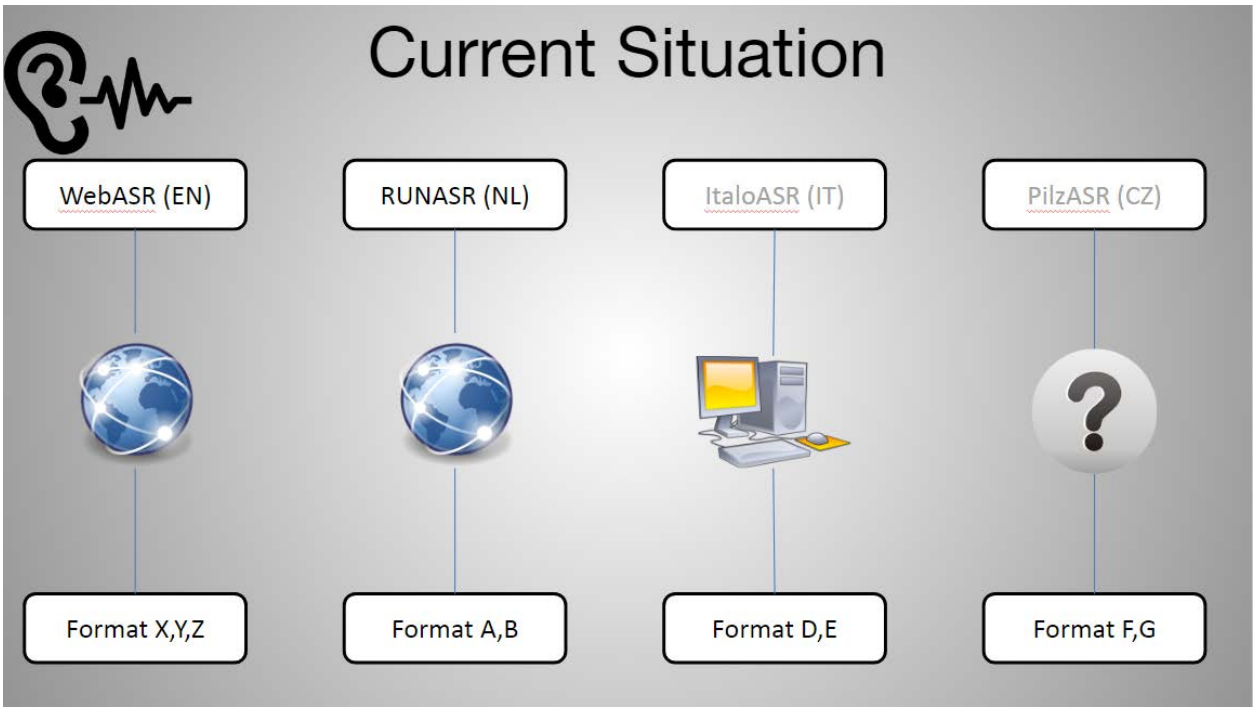


Figure 1: At this moment, each ASR-tool accepts just a few audio-formats, does not provide an audio-recoding into an accepted audio-format and exports the recognition results in their own format.

This should be transformed to a situation where there is one portal where the user can select a recognizer and can obtain the output text in a variety of mark-ups and formats:

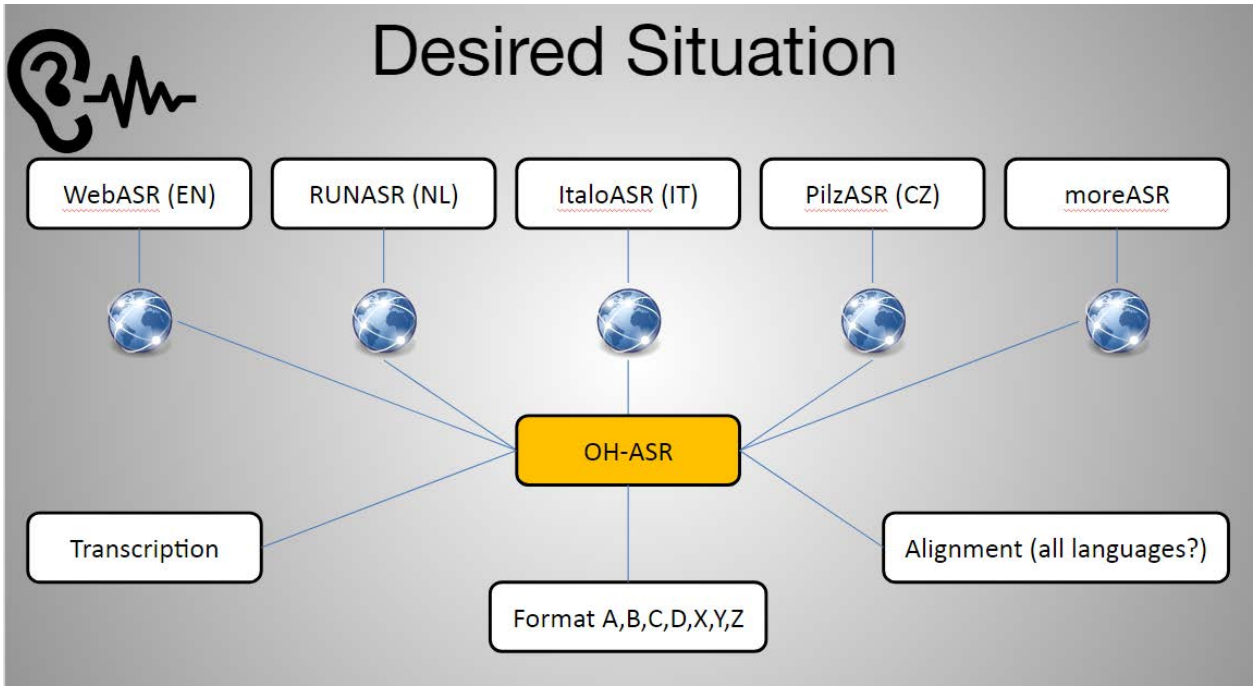


Figure 2: Once the portal is ready, recoding of the audio should be handled by the portal and the exported results will be in a wide variety of internationally accepted standards.

This basic idea will be detailed in a description of work with an elementary set of requirements. This description of work will be offered to a couple of knowledgeable parties that were preferably also present at the workshop. This should result in a proposal to carry out the work with 3 PM at max. The work should be completed this year.

Conclusion

Although there was a lot of comment, many questions and lots of suggestions, the Arezzo workshop can be considered as very successful. For the HLT-technicians it was a good opportunity to sit together, see what the others had done, discuss problems and solutions and further plans. For the proposed users, it was a good opportunity to see what is possible, how one can do it themselves and especially that the ASR-output becomes that good, that it can be used in their research/work.

Blogs

Some blogs, manuals and photos about the workshop, can be found here.

- Stef Scagliola: a blog about the workshop
- Arjan van Hessen: manuals and instructions about the tools mentioned in the workshop (work in progress)
- Henk, Silvia, and Louise: 3 short videos about "their" motivation for the Transcription Chain. All interviews are subtitled via ASR (Dutch, Italian and English) and automatically translated (Google Translate). It can be seen as a showcase of what is technically possible today. The 2 videos are used for DARIAH-teach as well.
- Photos: some nice images can be found on this Google Drive.