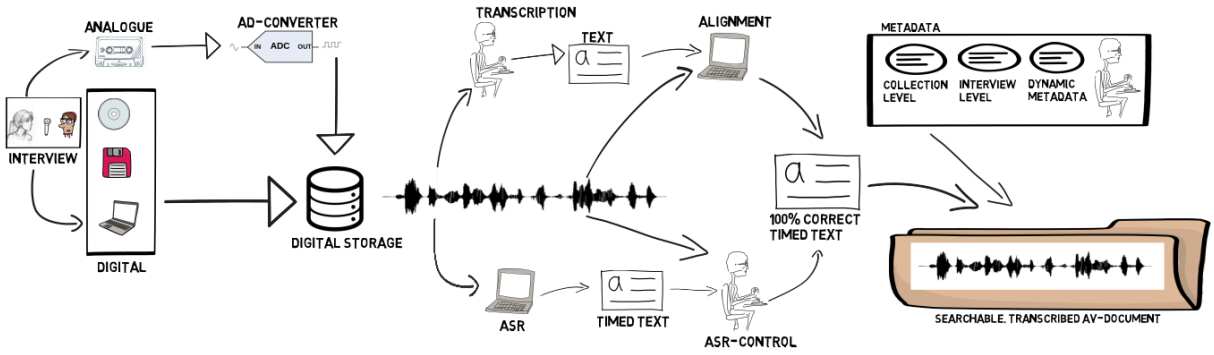
THE ORAL HISTORY CHAIN



A CLARIN PROPOSAL FOR THE REALISATION OF AN ***ORAL HISTORY TRANSCRIPTION PORTAL*** *USING EXISTING CLARIN - SERVICES SUCH AS:*

- *AUTOMATIC SPEECH RECOGNITION*
- *FORCED ALIGNMENT*
- *TRANSCRIPTION EDITING- AND CORRECTION TOOLS*

ONCE REALISED, THE PORTAL WILL BE ACCESSIBLE FOR ALL HUMANITY AND SOCIAL SCIENCE SCHOLARS FROM THE CLARIN COMMUNITY

HENK VAN DEN HEUVEL

ARJAN VAN HESSEN

STEF SCAGLIOLA

THE ORAL HISTORY CHAIN

A CLARIN-PROPOSAL

INTRODUCTION

The document below contains the requested information as specified in the CLARIN workshop II document and an overview of the workflow and time schedule. During the writing of this document, we felt the need to write an accompanying document, containing more in depth motivation about the choices made, the considerations and the potential problems envisioned. This document and the report about the Arezzo-workshop can be found on the website of this project: http://oralhistory.eu/workshops/ohtc-proposal

THE DEVELOPMENT TEAM

Information about the development team, including the those who are involved an advisory and/or monitoring role.

Two developers will work on the Oral History Transcription Chain. The first will (probably) be Sara Ahmadi; currently a PhD-student at the Radboud University who will have here PhD in September. Sarah was involved in the development of the current OH-recogniser of the CLST and is an outstanding programmer as well. She will be responsible for the OH-Transcription Portal V1.0.

The second developer will be someone (yet not none) from the group of Christoph Daxler who will integrate the WebMAUS and OCTRA part the portal.

The monitoring and advising of the developers will be done by Henk van den Heuvel, Christoph Daxler, Tomas Hain and Arjan van Hessen. Henk and Thomas will monitor the integration of "their" ASR-tools into the portal. Christoph will monitor the integration of the WebMAUS and OCTRA software into the portal and Arjan will monitor the overall design and access to manuals and help files.

COLLECTION OF USER REQUIREMENTS

Information on the approach towards the collection of user requirements

Most of the user requirements were collected before and during the workshop in Arezzo (see background document). However, we will publish various β-versions of the portal during the development and ask the Arezzo-scholars to comment on these versions. Before delivering the final version, more OH-scholars will be asked to try-out the portal and comment on functionality, understandability, user-friendliness and overall usefulness of the portal. Comments about missing items are welcomed but we cannot promise beforehand to include all these missing items in the first final version.

TRANSCRIPTION PORTAL

A summary of the results obtained, including performance figures, if available (max 400 words)

The result of the 3 months of development will be a Transcription Portal were (OH-)scholars can upload their interviews and download the automatically generated transcriptions (version 1). These timed transcriptions can be manually corrected and realigned via the integrated OCRA and WebMAUS functionality (version 1.1).

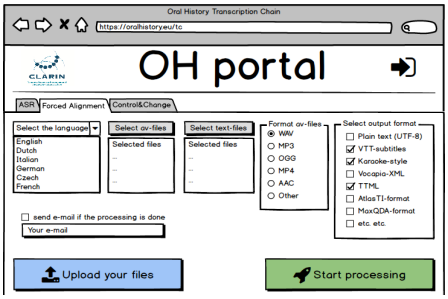


Figure 1: impression of the screen of the final Transcription Portal with some additional output filters (AtlasTI, MaxQDA).

INPUT FORMAT

The most commonly used audio formats (wav, mp3, ogg, flac, etc.) can be uploaded. If the audio-format used is too exotic, help can be asked via the help-button on the portal.

OUTPUT FORMAT

The output formats foreseen are: SRT, WEBVTT, Karaoke-html, and text. If it turns out in the evaluation that some formats are heavily missed, and there is some money/time, additional output formats may be considered.

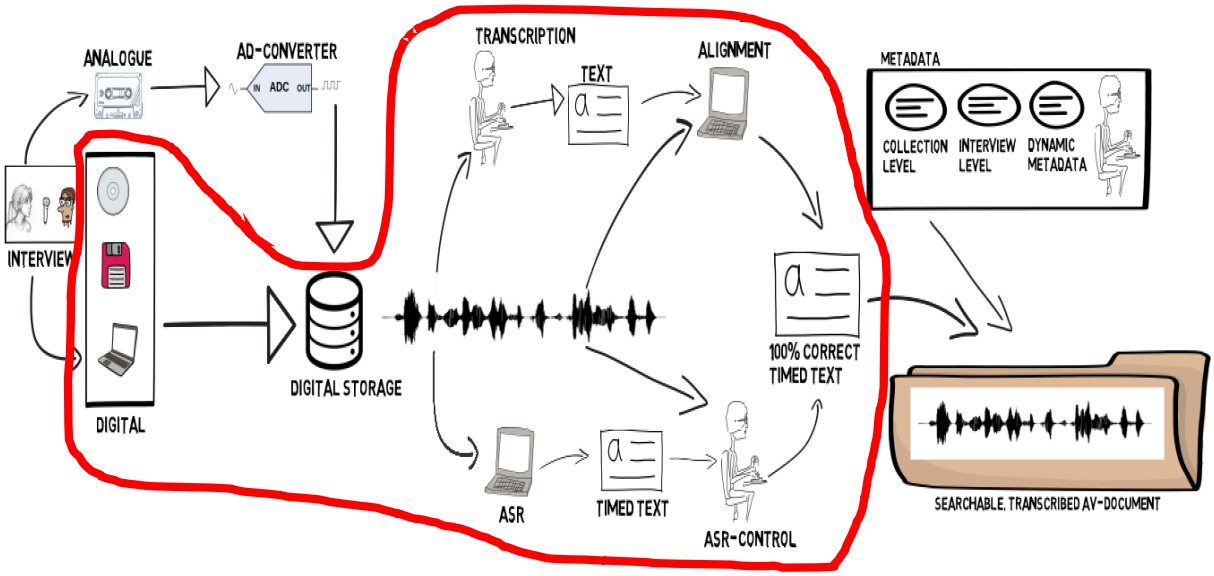


Figure 2: a graphic overview of the process "from interview to a document with transcriptions and metadata on the web. The transcription chain as proposed here, is the red-circled part. It contains the recoding, the recognition, correction and forced alignment.

STRATEGIC GOALS

Contribution of the workshop and project to strategic goals of CLARIN ERIC (max 150 words)
The Transcription Portal will be an open platform that enables the inclusion of:

a) additional ASR-engines, as long as the use the REST-protocol
b) additional output formats (for example the output on the phonetic level)

The open character of the portal will guarantee that additional functionality can be added. A portal like this Transcription Portal includes various technologies that are developed in and around the

CLARIN community. The portal will have a functional value (it works and scholars working with spoken narratives can use it for their work) and a more PR-value, showing that a combination of humanity scholars and computer/language-and-speech scientists can set up a technological service that is focused on the daily work of humanity scholars. In short: this is what CLARIN stands for.

INTEGRATION IN THE CLARIN INFRASTRUCTURE

Recommendations that should facilitate the integration in the CLARIN infrastructure, including suggestions for how the results could be brought under the attention of potential users (max 300 words)

Once realised, tested, and accepted, the Transcription Portal may become a part of the CLARIN services as published at: https://www.clarin.eu/portal

The portal will allow access from and to other software tools of the CLARIN infrastructure. This allows developers to send an audio-file from their program to the portal and download the results from the portal into their program. In this way, the portal is passed but the functionality is used. A good example of the integration of the functionality in the CLARIN infrastructure.

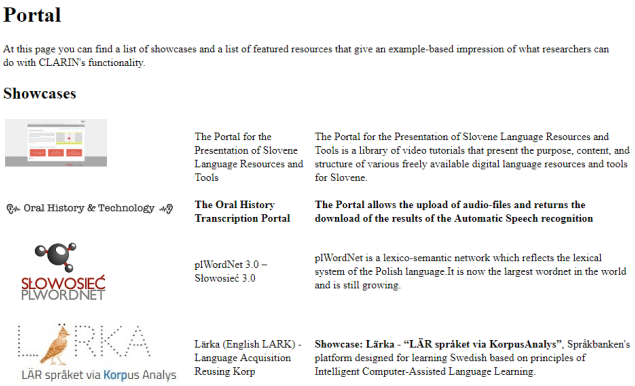


Figure 3: mock-up of a future CLARIN portal page that incudes the Transcription Portal.

PUBLICATIONS

Any publications planned

The portal will be showed on workshops, conferences (bazaar/demos) and other events where the targeted audience is present. Moreover, we will make screencasts (and place them on "our" websites) to help scholars using the portal.

POTENTIAL FOR IMPACT

A summary of any further potential for impact (max 200 words)

Once the portal is up-and-running, we will try to catch the attention of other potential ASR-providers who have ASR-engines for languages not available in the portal right now. Our first attempt will be to include ASR-engines we are aware of (Czech, French). Moreover, we will emphasis the fact that the Dutch, Kaldi-based, ASR-engine is available as open source on GitHub. This may be helpful for developers in countries that currently lack a good ASR-engine. If enough speech and language resources are available, it must be doable to build an ASR-engine within a year. This may have a significant impact on the use of HLT-tools in the CLARIN community.

DELIVERABLE, ALLOCATION OF RESOURCES AND PLANNING

To summarize what has been described in detail above: "the project proposes to build a website with the following selection-options: a button to select av-files, a button to upload av-files and a button to start the process. After clicking the submit button the audio files will be processed by the ASR-engine, and the recognition results will be available as files to be downloaded from the same screen.

In the Arezzo-workshop it became clear that this chain of tools is already performing quite well. However, the transcoding (of the AV-files and of the recognition-results) must currently be done by the end-user/scholar on his/her own computer.

Our estimate is that it is not very difficult to realise the TC V1.0: a basic service that encodes the audio, recognises the audio-files, encodes the output in the desired format(s) and sends the results to the user.
Our expectation is that we can realise V1.1 (integration of the Forced Alignment tool (WebMAUS) and the correction tool (OCTRA)) as well within the available 3 PMs.

The efforts with regard to the activities that will be initiated are distributed as follows:

1 PM: build portal with conversion tools
1 PM: include ASR servers for NL, EN (and ...)
1 PM: include webMAUS and OCTRA for version 1.1

COORDINATION

The project will be coordinated by RU (Henk van den Heuvel).

T +31 24 3611686
E H.vandenHeuvel@let.ru.nl
W http://www.ru.nl/english/people/heuvel-h-van-den/

WORKFLOW AND TIME SCHEDULE

Week #	*Activities*
1-2	Reading, assessing how the various REST-services work, writing scripts to ensure access the different ASR-engines and other software packages
3	Mock-up of the first version, asking feedback from the other stakeholders
4	Finalizing the requirements for version 1.0
5-7	Realizing TC-portal V1.0 ß. No (serious) layout but just functionality
8	Testing of version V1.0 by other stakeholders.
9	Incorporation of the comments of the potential users (the stakeholders)
10	Mock-up of the second version, a second round of feedback from community of scholars
11-12	Realizing TC-portal V1.1 ß. No (serious) layout but just functionality
13	Testing of version V1.1 by other stakeholders. beautification of the portal
14	Writing the final report, dissemination, describing future/desired additions

PARTICIPANTS

An overview of the 21 participants at the Arezzo-workshop. The people came 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