# Application for funding for a CLARIN workshop

| | |
|---|---|
| **Name workshop** | ORAL HISTORY:<br>Supporting the Transcription Chain |
| **Type** | II: Workshop with follow-up implementation project envisaged |
| **Proposers** | Henk van den Heuvel; CLST Radboud University<br>Arjan van Hessen; Utrecht University / University of Twente<br>Silvia Calamai; DSFUCI, University of Siena<br>Louise Corti; UK Data Archive, University of Essex<br>Stef Scagliola; University of Luxembourg<br>Martin Wynne; IT Services, University of Oxford |
| **Planned date** | November 2016 & March 2017 |
| **Budget** | €13,650.00 |
| **Summary** | As a follow up of the CLARIN workshop on Oral History (OH) archives in Oxford (April 2016), we apply for funding to prepare two workshops. The first workshop in November 2016 is a meeting in which the proposing team will prepare the second workshop. The second workshop is scheduled for March 2017. The two workshops applied for address the design and development of an (orthographic) transcription chain specifically designed for OH interviews including Analogue-to-Digital-conversion, automatic speech-to-text modules, and crowdsourced transcription platforms. The envisaged outcome of the two workshops is an implementation plan for a OH transcription chain that can be integrated in the CLARIN infrastructure. |
| **Contact person** | Dr. Henk vd Heuvel<br>E-mail: H.vandenHeuvel@let.ru.nl<br>Telephone: +31 24 3611686 |

## 1. Goal of workshops

This application addresses and the design and implementation of transcription chain for oral history (OH) interviews. Two workshops are needed. The initial workshop in November 2016 is a meeting in which the proposing team (hereafter the steering committee) and representatives of 2 or 3 CLARIN Centres will prepare the second workshop. The second workshop is scheduled for March 2017. (A third workshop addressing the workflow for ensuing analysis of transcribed oral documents will be the subject of a funding proposal later in 2017).

Central in the present application is the second workshop. Its goal is to draft a proposal for establishing an infrastructure directed at the orthographic transcription of existing OH-collections, through the integration of existing modules. This infrastructure contains three basic building blocks, all available in open source:

- **Analogue-to-digital (AD):**
  digitising analogue material in such a way that it resembles the original audio quality AND becomes optimal suitable for ASR.
- **Automatic Speech-to-Text (ASR):**
  Making a service where non-technical people easily can upload AV-recordings and retrieve the recognition results.
- **Transcription improvements:**
  Offering a web-based service where scholars can adjust the errors made by the ASR-engine e.g. in a crowdsourcing environment

The initial version of the proposed infrastructure will support 3 languages: English, Italian, and Dutch. But the openness of the infrastructure must guarantee that other languages can be added, if/once a good open source ASR-engine for that language is available.

Goal of the second workshop is to bring together scholars with expertise in automatic speech recognition & tools, crowdsourcing experts, representatives of the OH research field and CLARIN infrastructure experts with the following objectives:

- Exchanging information about the APIs of relevant components of the envisaged infrastructure and about the requirements of the end users and decide on which functionalities/variants to be integrated in the OH transcription chain
- Draft implementation plan for a versatile OH transcription chain that can be integrated in the CALRIN infrastructure involving 3 PM development effort.

## 2. Link with CLARIN's strategic priorities

The significance of Oral History in CLARIN's priority setting is underlined by a CLARIN-PLUS 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. The proposal resulting from the workshop proposed here should contain the detailed work plan to realise an optimal implementation of this transcription chain.

## 3. Profile of proposing team

The current proposal is a follow-up of one of the action lines from the CLARIN-PLUS Workshop on Oral History Archives[1] in Oxford in which all members of the proposing team participated.

Dr Henk van den Heuvel has been involved in the collection, compilation and validation of many spoken and written language resources at the national and international level. He has been project leader and project participant in CLARIN-NL projects amongst which the oral history INTER-VIEWS project (http://www.clarin.nl/node/267). He is also co-coordinator of CLARIN-NL's Data Curation Service.

Dr Arjan van Hessen. As a member of the CLARIAH-EB he is responsible for the user involvement: bringing humanities and computer science together and make things work. Moreover, as a HLT-specialist he is working in the field of OH since 2001; trying to implement HLT in the OH-community. In 2015 he was one of the main contributors of the Dutch NWO-groot proposal LISTEN (not granted).

Dr Silvia Calamai is Associate Professor in Linguistics at the University of Siena and she is the scientific co-coordinator of the Project on Oral archives *Grammo-foni*[2] *(Gra.fo) "Le soffitte della voce"*. She is associate member of IPinCH The Intellectual Property Issues in Cultural Heritage project (Simon Fraser University, Canada) and member of the Dariah-FR network on "Legal and ethical issues in digital research".

Louise Corti is an Associate Director at the UK Data Archive, since 2000, and currently leads the UK Data Service functional areas of Collections Development and Producer Relations. Her current research activities are focused around standards and technologies for ingesting, archiving and presenting digital social science data, particularly using open source infrastructures and tools. She has published widely on research data management for the social sciences. Before joining the Archive in 2000, Louise

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

[2] http://grafo.sns.it

helped establish, as Deputy Director of Qualidata, the world's first national qualitative data archive, from 1994. http://data-archive.ac.uk/about/staff?sid=corti

Dr Stef Scagiola is a postdoc researcher. She was affiliated to the Erasmus University working on the projects AXES, CroMe and Oral History Today. Since July 1, 2016 she is working at the University of Luxembourg as a historian, specialized in digital audiovisual archives, with an emphasis on oral history collections. From 2006-2011 she was the coordinator of an oral history project conducted at the Netherlands Institute for Veterans which resulted in the first large scale 'digital born' interview collection in the Netherlands. It consists of more than 1000 life-history interviews from among a representative sample of Dutch war- and military mission veterans. http://www.eur.nl/erasmusstudio/people/current_members/dr_si_scagliola_stef/

Dr Martin Wynne is Senior Research Support Officer at the University of Oxford. Martin is responsible for the Oxford Text Archive, which also involves managing the distribution of the British National Corpus (BNC). He currently has roles in the Oxford e-Research Centre, where he works as part of the Digital Humanities at Oxford initiative. Martin is also a member of the Faculty of Linguistics, Philology and Phonetics. Martin is a Director for User Involvement for CLARIN ERIC, and he organised the CLARIN-PLUS workshop on Oral History in Oxford, 18-19 April 2016.

## 4. Indication of profile(s) and number of envisaged participants

The preparatory workshop in November 2016 will be held with the members of the proposing team (6 people) and 4 representatives of CLARIN Centres willing/interested to host the envisioned services.

For the second workshop we would like to bring together about 15 to 20 experts in the following fields:

- Audio and speech processing (tools AD conversion)
- Automatic Speech Recognition (tools ASR)
- Crowdsourcing for the humanities (crowdsourcing platforms)
- CLARIN (tools infrastructure)
- Oral history research (envisaged end users of the tools)

The participant list of the CLARIN PLUS Workshop in Oxford has provided a rich palette of experts and will be used as starting point for the invitations.

The number of max 20 participants is considered well sized for intensive discussion and dedicated workplan writing.

## 5. Provisional information on workshop date(s), location and timetable

| Location | Time | Details/outcomes |
|---|---|---|
| Utrecht, NL | Second part of November | 1,5 days (afternoon, evening, morning) |
| Writing period, Internet | December - January - February | A set of documents in GoogleDoc and a website with, examples and showcases that should feed the design of the transcription chain will be set up and shared with the participants of the second workshop. |
| Arezzo, IT (DSFUCI, University of Sienna) | March 2017 | 2,5 days. The result of the second workshop is an implementation plan, based on a design informed by end-user requirements and expertise from the developers of the components to be integrated, plus identification of 2 or 3 CLARIN centres that will host the envisaged service. |

**Timetable:**

| Date | Activity |
|---|---|
| November 2016 | First workshop: team meeting in Utrecht, the Netherlands |
| November 2016 | Send out invitations, arrange details of second workshop, travel information, venue, programme, presentations slots |
| November - December | Workshop participants to contribute to the preparatory document with the various design options and user requirements for the transcription chain |
| January 2017 | Finalise organisation of the workshop, and finalise guidelines for presentations |
| March 2017 | Second workshop in Arezzo |
| April 2017 | Finalisation of implementation plan |
| *May 2017* | *Proposal for third workshop on data analysis, with select oral history data resources and tools hosted by the University of Essex on their big data platform, Colchester.* |

## 6. Provisional agenda Second workshop

Second workshop in Arezzo in March 2017:

| Day 1 | Information exchange, exploring the options and alternatives |
|---|---|
| Morning | travel time |
| 14:00 | Welcome |
| 14:15 | Overview of the workshop<br>General overview of the envisaged transcription chain<br>Presentations about suitable AD-conversion (tools)<br>Presentations about ASR (tools)<br>Design session: connecting the building blocks<br>Presentation of the skeleton of the service by the developer |
| 19:00 | Dinner |

| Day 2 | Information exchange, settling the building blocks |
|---|---|
| 9:15 | Summary of day 1 and overview of day 2<br>Presentations about manual transcription correction services<br>Presentations about Crowdsourcing strategies and platforms<br>Presentations about requirements OH researchers<br>Discussion |
| 13:00 | Lunch |
| 14:00 | Group discussion per building block resulting in a reasoned advice for each block<br>Plenary discussion on the advices and conclusions<br>Establish working groups for various parts of the proposal<br>Definition of the requirements (necessary and nice-to-have) of the service(s) to build and a discussion about the feasibility of the various components |
| 19:00 | Dinner |

| Day 3 | Proposal preparation |
|---|---|
| 9:15 | Group activities: drafting implementation plan<br>Plenary: concluding actions for finalising the implementation plan<br>Setup of the time schedules for the next months |
| 13:00 | Lunch |
| 14:00 | Adjourn |

## 7. Summary of envisaged implementation project

The proposed implementation will be based on the design agreed in the second workshop To appoint the various responsibilities a light management structure for the project will be set up.

Based on the implementation plan, a developer (N.N.) will write a more technical framework of the service and implement a first version of the proposed service to be built. This first version will serve as a demonstration service (one language, limited documentation, basic interface) for the second workshop. The writing and implementation will be done in the period between the first and second workshop; the estimated amount of time will be one month.

After the second workshop, the developer will implement the final version, based on the provided requirements of the second workshop. The estimated amount of time will be two months.

### Audio conversion

The great variety of AV-resources will guarantee a high number of different AV-formats. A small service will be made that "transforms" the AV-format into a format accepted by the ASR-engine. Many online ASR-services do have such a service (VOCAPIA, SPRAAK, HavenOnDemand), so it is expected that it will be an easy task to do.

### Speech recognition

With the availability of KALDI, a good, deep-learning based ASR-engines, became available for the academic community. In various countries "local" KALDI-based ASR-engines are currently developed. An example is the Dutch KALDI-based ASR engine developed by dr. Laurens van der Werff in cooperation with the University of Twente, the NISV, the Dutch Police, the FIOD (Dutch tax authorities) and Telecats (a Dutch SME).

### Language models

An important part of the ASR-engine is the Language Model (LM): the statistical model that predicts the change of word X, given a set of N previous spoken words. Dedicated LM are essential for interviews that talk about particular topics, using non-common or rarely used words that are typical for those topics. An interview in Italian about concentration camps will probably contain German words such as Führer, Kapo, Nacht-und-Nebel, Zyklon B, etc.

Building a LM is complex, but suitable open source software to build, is available.

### Transcription Corrections

Despite the high quality of modern ASR-engine, there will be (too much) errors in the transcription. If the transcription is only used for searching in the AV-recordings, a Word Error Rate (WER) of 30% is acceptable, but if it will be used for reading, subtitling, (automatic) translation or dialogue analysis, a nearly perfect transcription is necessary (WER<4%).

A web based service needs to be built in order to offer the users (the OH-scholars) the possibility to correct the recognition errors. Open Source initiatives to improve transcriptions and to automatically translate them do exist (Subtitle Edit Online) and can probably be integrated into an international platform such as Zooniverse. A crowdsourcing structure (for example Crowdflower) will be setup to facilitate the transcription corrections. The infrastructure must contain the possibility to integrate such services in the workflow. The integration can be via an inclusion of the software and/or via an export-import option.

### Export

The infrastructure must contain the possibility to export the transcriptions into formats that can be read by others, often used annotation tools used for analysis of oral history, such as ATLAS-TI, TEI and others.

### Maintenance and IPR

An important part of each infrastructure project is the maintenance. Who is responsible for the service, where will it run, is there enough money to run the service and update it when necessary for the next 5 to 10 years? What governance rules need to be in place if data have confidentiality issues?

CLARIN offers a sustainable infrastructure for the services described above, but robust agreements need to be made.

Building Language Models may infringe National Intellectual Property Rights. The steering committee will discuss this with the CLARIN Legal Issues Committee.

### Promotion, communication and training

It is important that tools and services are widely promoted across different user bases, spanning the disciplines who are interested in oral history and other spoken word resources. The second workshop will devise strategies and ideas for rolling out promotional, guidance and training materials to facilitate uptake on a national and international level.

## 8. Budget breakdown

| **Utrecht workshop** | **Persons** | **Days** | **Cost** | **Total** |
|---|---|---|---|---|
| Participants | 10 | 1 | | |
| Travel costs (F, 6 persons, € 300) | 6 | 1 | € 290,00 | € 1.740,00 |
| Travel costs (L, 4 persons, € 50) | 4 | 1 | € 25,00 | € 100,00 |
| Accommodation (€120) | 7 | 1 | € 120,00 | € 840,00 |
| Breaks & Lunch (€20) | 10 | 1 | € 20,00 | € 200,00 |
| Diner (€45) | 10 | 1 | € 45,00 | € 450,00 |
| Venue costs (none) | 1 | 1 | € 0,00 | € 0,00 |
| **Total** | | | | **€ 3.330,00** |

| **Arezzo workshop** | **Persons** | **Days** | **Cost** | **Total** |
|---|---|---|---|---|
| Participants | 20 | 2 | | |
| Travel costs (F, 16 persons, € 280) | 16 | 1 | € 280,00 | € 4.480,00 |
| Travel costs (L, 4 persons, € 50) | 4 | 1 | € 25,00 | € 100,00 |
| Accommodation (F, €90) | 18 | 2 | € 90,00 | € 3.240,00 |
| Breaks & Lunch (€20) | 20 | 2 | € 20,00 | € 800,00 |
| Diner (€ 35) | 20 | 2 | € 35,00 | € 1.400,00 |
| Venue costs (€150) | 1 | 2 | € 150,00 | € 300,00 |
| **Total** | | | | **€ 10.320,00** |

| **Preparations for the workshop(s) & finalisation of the implementation plan** | | **Hours** | **Rate** | **Total** |
|---|---|---|---|---|
| Average 20h/pp a €55 | 10 | 20 | € 55,00 | € 11.000,00 |
| of which in kind | 10 | 20 | -€ 55,00 | -€ 11.000,00 |
| **Total** | | | | **€ 0,00** |

**Total** **€ 13.650,00**

*Legend: F = Foreign participants, L = Local participants*