Application for funding for a CLARIN workshop

Name workshop	ORAL HISTORY: USERS AND THEIR SCHOLARLY PRACTICES IN A MULTIDISCIPLINARY WORLD: EXPLORING HOW USERS WORK WITH TOOLS ON ORAL HISTORY DATA
Type	I: Workshop aiming to meet CLARIN strategic priorities
Proposers	Louise Corti; University of Essex Stef Scagliola; University of Luxembourg Silvia Calamai; DSFUCI, University of Siena Henk van den Heuvel; CLST Radboud University Arjan van Hessen; Utrecht University / University of Twente Christoph Draxler; Ludwig-Maximilians-Universität München Institute of Phonetics and Speech Martin Wynne; University of Oxford
Planned date	19-21 September 2018
Budget	€ 16.500,00
Summary	The proposed Type 1 workshop follows a stream of concerted activity around the exploitation of techniques and tools for working with oral history (OH) data. This workshop focuses on the analysis phase of the research process, building on the work already done to bring together CLARIN technologies for speech retrieval and alignment, 'Transcription Chain'. The workshop will bring together scholars to explore the diversity of scholarly practices across CLARIN-relevant disciplines - digital humanities, OH, linguistics and traditional social science - in using tools (CLARIN or otherwise) to interrogate and analyse OH data sources. The workshop will elucidate how CLARIN can better support these diversity of practices and lower barriers to the use and take up of its resource and technologies.
Contact person	Louise Corti, Uk Data Archive, University of Essex, UK. corti@essex.ac.uk tel: +44 1206872145

1 Goal of workshop

The workshop will bring together users from contrasting CLARIN-relevant disciplines who use OH sources in their daily work - digital humanities, linguistics, oral history and traditional social science. It aims to elucidate their everyday research practices and how they might be best encouraged to take up CLARIN tools in order to expand the community of researchers that can benefit from CLARIN. As well as crossing language borders, it also offers an opportunity for crossovers with data infrastructures that are evolving in the domain of the social science, such as CESSDA.

The key objectives are:

- exploring OH scholars' needs and practices around the use of tools and techniques for processing, analysing or publishing OF data;

-

- broadening the CLARIN user base and tools portfolio;
- opening up cross-disciplinary dialogue (such as fostering CESSDA collaboration) and a space for open-source software and promotion of CLARIN tools,including gaining currently untapped insights from social scientists;
- introduce CLARIN DH approaches and tools e.g. TChain and textual annotation tools to social scientists, and gauge reactions in a cross disciplinary setting;
- introduce OH and spoken word data resources to linguists as a new source: audio and text; and to tool developers.

2 Link with CLARIN's strategic priorities

This workshop builds on a stream of concerted activity in 2016-17 around the use of techniques and tools by researchers for working with OH data: the Oxford CLARIN-PLUS workshop and two CLARIN workshops in Utrecht and Arezzo, that focused on supporting the transcription phase of the research process. This workshop meets the following CLARIN strategic priorities:

- *User Involvement*: it brings together scholars with different practices to explore and offer feedback on CLARIN language resources and tools, which contributes to reaching out beyond the usual DH communities;
- *Legal issues and copyright*: sample data resources will be created as FAIR open data, that can be used for future interrogation by CLARIN projects;
- *Crossing borders*: The workshop brings together language resources and scholars as participants from 4 languages (English, Dutch, German and Italian) across academic, data archiving, public library sectors and industry (e.g. analysis software provision). The aim is to cross borders of both country and discipline, and to foster closer collaboration with other key research infrastructures, like CESSDA and DARIAH.

3 Profile of proposing team

Louise Corti is an Associate Director at the UK Data Archive, since 2000, and currently leads the UK Data Service Collections Development and Data Publishing teams. 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 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 Scagliola is a postdoc researcher. She was affiliated to the Erasmus University working on the projects AXES, CroMe and Oral History Today. At present she works at the Luxembourg Centre for Contemporary and Digital History, where she is developed an online teaching platform on Digital Source Criticism. From 2006-2011 she coordinated an OH project at the Dutch Veterans Institute resulting in a collection of 1000 life-history interviews with Dutch war- and military mission veterans. http://www.eur.nl/erasmusstudio/people/current_members/dr_si_scagliola_stef/.

Dr Silvia Calamai is Associate Professor in Linguistics at the University of Siena and scientific co-coordinator of the Project on Oral archives *Grammo-foni (Gra.fo) "Le soffitte della voce"*. She is member of the CLARIN Legal Issues Committee, of the board of the Italian Speech Sciences Association (AISV), and of the scientific committee of the Historical Archive of Arezzo's psychiatric hospital. At present, she coordinates the project *Chinese Culture and Languages in Italy* (Wenzhou University & Siena University).

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 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 OH INTER-VIEWS project (http://www.clarin.nl/node/267). He is also co-coordinator of CLARIN-NL's Data Curation Service.

Dr Christoph Draxler is a researcher at the Phonetics Institute, University of Munich, since 1991. He is project lead of CLARIN-D at BAS and software architect and developer for a number of speech tools including SpeechRecorder and SpeechFinder. He also teaches and is a thesis supervisor.

Martin Wynne is active in the Digital Humanities team in OeRC. He is UK National Coordinator for CLARIN, a major pan-European initiative to build a research infrastructure for the creation and use of electronic language resources in the Humanities and Social Sciences. He was also on the executive committee and board of directors of CLARIN. As national coordinator, he leads the CLARIN-UK consortium of research centres with an interest in using, developing and sharing digital language resources and tools.

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

The proposing team and participants of the past 3 OH-focussed workshops have created a community of experts (CLARIN and non-CLARIN) from The Netherlands, Great Britain and Italy who have actively scoped and assembled invitations to like-minded scholars from other communities of practice. These countries have been chosen on the basis of the availability of mature open source speech retrieval software.

The workshop, hosted by a CLARIN Centre in Munich, will invite around 25 Dutch, British and Italian and German scholars and archivists who work /teach with oral history data and share a proactive methodological interest in exploring how technology can support new ways of annotation and analysis. They represent the following communities:

- Digital historians and social science users who undertake research analysing OH data sources;
- Linguists who use spoken language sources;
- Tools specialists (CLARIN and otherwise) who develop support data analysis tools.

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

Location	Time	Details/outcomes
Munich, DE (Institute of Phonetics,	19-21 Septemb er 2018	2.5 days. The outcomes of the workshop is a report on user needs for tools for oral history research, seeking to shed better knowledge on the different techniques and tools currently used by digital humanities and social science

University of Munich)		scholars and might be needed by CLARIN to bridge gaps to support infrastructure and tools.

Timetable:

Date	Activity
October 2017 – March 2018	Web based team planning meetings
March 2018	Send out invitations, confirm details of workshop, travel information, venue, programme, presentations slots and venue/ IT checks
March - July 2018	Assemble open source digital data language resources by common theme in 4 languages, Ensure that it is processed in ways that make it suitable to analyse with CLARIN tools
July - Septembre 2018	Workshop participants to contribute to the preparatory document to scope tools and OH sources currently used, and their experience with other related tools
August 2018	Finalise organisation of the workshop, and finalise guidelines for presentations
September 2018	Type I Workshop in Munich
September - October 2018 September - November 2018	Write brief report based on User feedback: a. A matrix in which the experiences of scholars from different disciplines with the various tools that are worked with during the workshop are documented b. Recommandations to CLARIN based on this feedback and presented at the CLARIN Conference Pisa in October Write a methodologically--oriented publication arising from the workshop, comparing linguistic approaches to working with oral history data
September - October 2018	A series of short video clips with comments by contributors on what they have learned and what has surprised them. To be published on the Oralhistory-eu and CLARIN website. (CLARIN support would be needed for this work)
October 2018	Presentation at CLARIN Conference, Pisa
October - December 2018	Formulate an agenda for the future.

6 Provisional agenda workshop

Munich, 2 full days (Weds lunch to Friday lunch) 19-21 September 2018:

Day 1	Information exchange, exploring the options and alternatives
Morning	travel time
14:00	Welcome
14:15	Overview of the workshop General overview of CLARIN mission and tools: - what computational linguists do with spoken corpora - what oral historians do with interviews - what social scientists do with interviews - envisioned Digital Humanities approaches to oral history data - can historians, linguists and social scientists share tools? - Introduction to the selected OH and spoken word resources. These will be centrally mounted and users will have option to download them to their laptops. - explanation of the workshop tasks - short demonstration of the TChain workflow
19:00	Dinner

Day 2	Information exchange: envisaging and embracing others techniques and tools

-

9:15	Assemble into language groups Testing the TChain: - try out of Tchain with Dutch, English, German and Italian data, or participants' own data - document experiences - evaluation Presenting annotation tools - ELAN and proprietary tools - try out different annotation tools with your own data or with shared data - what are the differences? - are there computational elements included? - how are sound and moving image included? - document experiences - evaluation
-	Lunch
14:00	Presenting linguistic preprocessing and concordance tools - try out with linguistic tool 1 on oral history data that has been preprocessed (GB, I, NL, G)* - how do linguistic research questions relate to the social science and history paradigms? - when does scale matter to be meaningful? - document experiencers
19:00	Dinner

Day 3	
9:15	Presenting linguistic preprocessing and concordance tools - try out linguistic tool 2 on oral history data that has been preprocessed * - what kind of meaning can be extracted from this? - what is the difference with the first tool? - document experiencers - evaluation of tool 1 and tool 2 Plenary discussion on conclusions and requirements for OH researchers and interdisciplinary opportunities in the CLARIN infrastructure * Scholars can bring in their own data only if they have sent it in advance in order to be preprocessed.
13:00	Lunch
14:00	Adjourn

7 Envisioned outcomes of the workshop

1. A short report based on a matrix documenting the experiences with the transcription, annotation and linguistic analysis tools by scholars differentiating in discipline; including a presentation and poster at the CLARIN Conference in Pisa on 8 October.

2. A methodologically-oriented publication and presentations arising from the workshop, comparing approaches to working with oral history data.

3. An agenda for the future that fosters a bigger network of those who care about oral history data, and brings together digital humanities and social scientists and oral historians. This could have the form of a continued network that could offer input in CLARIN work plans in a structured way through a paid coordinator.

-

4. A series of short video clips with comments by contributors on what they have learned and what has surprised them, that will be published on the Oralhistory-eu website and on the CLARIN website.

Appendix: relevant CLARIN and non-CLARIN tools

It is important to clarify what kind of CLARIN tools used by linguists and tools used by social scientists and oral historians for exploring spoken word data will be introduced to the meeting. The first will be an assessment of the Transcription Chain as a data preparation tool, second some of the current spoken and textual annotation tools available, the third will be classical tools used by linguists. We have included presenters who have expertise in these tools,and would value the input from the central CLARIN office who has the best overview of tools and user needs, e.g. Darja Fiser. While the workshop will focus on audio-material, we will also consult briefly on use of annotation of visual traits, and for emotion recognition of audio, bodily movements.

A. Transcription Chain (CLARIN)

The TChain web portal being constructed and aiming to complete by the time of the workshop will consist of an interactive website with the following selection options:

- A button to select audio files
- A button to upload audio files
- A button to start the process (submit).
- 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 various formats, with and without time-stamps;
- ASR engines for OH material will be provided for
    - English (Sheffield)
    - Dutch (RU Nijmegen)
    - if possible, also for German, Italian and Czech.
- Hopefully, integration of the Forced Alignment tool (WebMAUS) and the correction tool (OCTRA) in the TChain webportal.

B. Text and audio annotation tools

The tools listed below will be reviewed by the team in advance of the event to pick out the most salient for the workshop. An important distinction for non linguists is the difference between computational annotation, and manual annotation.

Corpus Workbench (CWB) online service

CWB (non-CLARIN) is a corpus analysis engine underlying a number of services, including CQPWEB, CWB can be used for online analysis of corpora. Texts need to be preprocessed and converted to a specific format and then need to be ingested by a site administrator. CQPweb at Lancaster is an example of an online interface built on the Corpus Workbench.

CLAWS and USAS taggers

These are wordclass semantic taggers (non-CLARIN) for modern British English, available for free. Adding these annotations to texts helps with users to search and explore the texts, making it possible

to find all forms of a word (e.g. 'torture', 'torturing', 'tortured', etc.), to disambiguate certain words ('meeting' as a noun or verb, 'shoot' as a term in warfare or photography), and to detect patterns of usage ('there are lots of words relating to food and drink in this text'). These are widely-used applications, not currently integrated into the CLARIN infrastructure, and should be considered as candidates for integration.

AntConc to download

AntCon (non-CLARIN) is a free, easy to deploy, multi-platform and easy to use, Antconc offers a range of functions to explore and analyse texts and language corpora. The functions are designed by and for linguistics, and include concordance, collocation, clusters, word frequency lists and keywords. The suite of software applications also include TagAnt (a deployment of the Treetagger) These are a widely-used set of applications, not currently integrated into the CLARIN infrastructure, and should be considered as candidates for integration.

GATE online

GATE (non-CLARIN) is a suite of applications for natural language processing, including annotation and analysis functions, e.g. named entity recognition. A widely-used set of applications, not currently integrated into the CLARIN infrastructure, and should be considered as candidates for integration.

Natural Language Toolkit (NLTK) online

NLTK (non-CLARIN) is a widely-used suite of applications for natural language processing. They are not currently integrated into the CLARIN infrastructure, and could be considered as a candidate for integration.

Treetagger online

Chunker (CLARIN) is a wordclass tagger and constituency parser for English, Dutch, Italian and German. Can be deployed as a standalone application, or via Weblicht.

ELAN download

ELAN is a professional tool for the creation of complex annotations on video and audio resources, widely used by language and speech scholars. Users can add annotations to audio and/or video streams, such as a sentence, word or gloss, a comment, translation or a description of any feature observed in the media. Annotations can be created on multiple layers, called tiers, and can be time-aligned to the media or refer to other existing annotations. The textual content of annotations is in Unicode and the transcription is stored in an XML format. This is not used by social scientists but could be very useful.

WebLicht online

WebLicht (CLARIN) is an orchestration engine for web-based linguistic annotation tools, distributed repositories for storing and retrieving information about the tools, and this web application, which allows you to easily create and execute tool chains without downloading or installing any software on your local computer. This application and its associated tools are continually being updated and improved. Applications available via Weblicht can be found in the Virtual Language Observatory.

NVivo download proprietary

NVivo (non-CLARIN, commercial) is a popular qualitative data analysis computer software package produced by QSR International. It has been designed for qualitative researchers working with very rich text-based and/or multimedia. It allows coding and annotation.

-

Atlas ti download proprietary

ATLAS.ti (non-CLARIN, commercial) is another popular tool to help researchers systematically analyze unstructured data (text, multimedia, geospatial). The program provides tools that enable the user to locate, code, and annotate findings in primary sources and to visualize the relations between them.

C. Text corpora

A number of corpora of spoken language are available, and which might offer the possibility to act as reference corpora, allowing comparison of oral history data with evidence of other speech events. This could allow the possibility to ask questions like “How does oral history data different to other types of speech?”. The British National Corpus (http://purl.ox.ac.uk/ota/2554), and Spoken BNC 2014 are general reference corpora for contemporary spoken British English. The BNC is discoverable via the VLO, but are otherwise not currently integrated into the CLARIN infrastructure.