London Interdisciplinary Social Science Doctoral Training Partnership

Advanced Research Methods in Social Sciences

Course Code & Title	**LISS293 Designing a Spoken Corpus: Data Collection and Transcription**				
Convenor(s)	Dr Robbie Love				
Institution	Aston University		**Department**	LISS DTP	
Academic Year	2021-22		**Term**	Spring	
Number of sessions	3	**Research Platform**	• Digital & Archival Research Methods (DAR)	**Length of Session(s)**	3 hours x 3
Day, Date			**Start : End**	**Room Location**	
Tuesday, 1st February 2022 Wednesday, 2nd February 2022 Thursday, 3rd February 2022			14:00-17:00	Via Zoom	
Enrolment Links:	Click here to register for this course on Skillsforge.				

Course Description:

A corpus is a large, planned collection of 'real life' examples of language. Corpora are widely used to influence and inform areas such as language teaching, dictionaries, and new technologies (e.g. speech recognition software). They are also used by a range of researchers interested in languages, social sciences, and humanities. This module explores the important early stages of designing a corpus for language analysis.

When working with audio recordings such as oral histories or interviews, there are many issues to consider around creating precise and useful transcriptions to populate the corpus. These include: accurately reflecting what was recorded; capturing key details about the speakers, recordings, and contexts; and organisation and coding for current and future analysis. As well as examining these areas, there will be the opportunity to discuss and work on your own (planned or actual) data collections.

A series of 3 workshops will be run by Dr Robbie Love, who worked on the British National Corpus 2014 at Lancaster University.

The course will include:

1. Metadata and coding files - determining categorisations, deciding levels of detail to record for:
    i. speaker(s) information
    ii. recording information
2. Dealing with non-standard variants in transcription e.g. dialect pronunciations
3. Inclusion of non-lexical items e.g. pauses, exclamations
4. Recording non-speech information e.g. external noises, contextual events
5. Protecting anonymity
6. XML and TEI specifications and standards of compatibility
7. Overview and evaluation of software options - including automatic and time-aligned transcriptions

London Interdisciplinary Social Science Doctoral Training Partnership

Advanced Research Methods in Social Sciences

Reading List:

Before attending please read the following two chapters:

Love, R., 2020. *Overcoming challenges in corpus construction: The Spoken British National Corpus 2014*. Routledge.

Chapter 3: Theoretical Challenges in Corpus Design

Chapter 5: Challenges in Transcription Part 1 – Conventions

Further recommended reading lists will be provided before / during the sessions.

Eligibility:

Open to all PhD students in social science and humanities, most relevant to linguistics students. Participants should have some prior knowledge of corpus linguistics.

Pre-course preparation:

If you are working on your own data, or starting to collect your data – consider any questions you have about the transcription, and the planned design for your corpus. There will be opportunities throughout the sessions to discuss these.

Number of students:

Minimum number required to run: 2

Maximum number of places available: n/a