Data Curation Profile for Linguistics

Researcher: Dr Bert Remijsen, School of Philosophy, Psychology & Language Sciences, University of Edinburgh

Project: Metre and Melody in Dinka Speech and Song

Project dates: 2009 – 1012

Profile author: This profile was compiled by Anne Donnelly, Liaison Librarian, School of Philosophy, Psychology & Language Sciences, following an interview with Dr Remijsen on 20th August 2013.

1. OVERVIEW

1.1 Overview of the research project

The songs in this collection were recorded and annotated by researchers from the University of Edinburgh and the School of Oriental and African Studies (SOAS) in London. The project aimed to understand the interplay between traditional Dinka musical forms and the Dinka language (which distinguishes words not just by different consonants and vowels but also by means of rhythm, pitch and voice quality), and to learn more about the way the song tradition responded to the disruptions of the long Sudanese civil war. In this context, we aimed to record a large collection of Dinka songs for preservation in a long-term sound archive. This collection is the result of that effort. It presents song material from 36 Dinka singers and groups of singers.

1.2 Funding source

The Arts & Humanities Research Council's 'Beyond Text' programme.

1.3 Data management planning for the project

There was a data management plan for the project at the outset, as required as part of the submission for funding, although Dr Remijsen anticipated that this might change during the life of the project.

1.4 Overview of the data and research records related to the project

The dataset comprises 185 sound files in .wav format. A naming template was created at the outset that captured such details as the name of the singer, the songs record, the dates of the recording and the chunks of songs recorded. One song only, although included in the project records, was not included in the collection, as permission to disseminate it was withheld by the singer.

The sounds files have associated records comprising singer-specific permissions and questionnaire documentation, metadata for each song covering instrumentation, details of the recordings and a brief summary. All the information about both singers and songs is summarised in an index file which, together with a readme file, makes the data accessible to those interesting in using them.

1.5 Intellectual property owner of the data

Essentially the singers of the songs, although there were difficulties in ascertaining this at times. Permission to disseminate was withheld in one case because the song was not the singer's own. However, with this exception, the School of PPLS obtained permission to archive and freely disseminate the recordings on a 'not for profit' basis.

1.6 Approximate number of data files generated during the course of the project

185 .wav sound files – the songs - with related annotations, permissions and questionnaires.

1.7 The average size of the data files

The median file size was given as 22 Mb.

2. ORGANISATION

2.1 Data formats

The files are in .wav format.

2.2 Organisation of the data

The files are structured using singer, date of recording, song recording chunks.

2.3 Importance of the metadata system used for the project

Dr Remijsen considered the system used to have been a good and beneficial to the project; it employed a file naming convention that also enabled the identification of the file content without the necessity for an additional key.

2.4 Software programs and tools used in the collection and organization of this data

A shareware program - *Praat: doing phonetics by computer* – that facilitated varying levels of timeline annotations for each recording.

2.5 Software programs or tools required to utilize this data

Pratt: doing phonetics by computer. However, the .wav files are readily opened in either PC or Mac. Although the text grid and timeline annotations are in a software specific, this is not proprietary; *Praat* is shareware and thus its text grids can be successfully opened in other annotation software, e.g. ELAN, which is also widely used. Many of the Dinka song text grids have already been converted to a more widely available format.

2.6 Storage and back-up of files

During the life of the project the files were stored on a directory on the Linguistics & English Language (LEL) server within the School of PPLS, accessible to all project team members. Dr Remijsen took sole charge of and responsibility for the weekly updates, thereby ensuring

that all the materials were there while avoiding duplicate versions. The files were backed up automatically each night.

2.7 Measures to control access to the data

During the course of the project the directory on the LEL server was accessible through the internet, with the folder password-protected and accessible to project team members only.

3. STORAGE & SHARING

3.1 Future preservation of the data

The song collection files are now archived in Edinburgh DataShare for future management and sharing. They are also in the Max Planck Institute in Nijmegen in the Netherlands, which documents endangered languages. This collection also feeds into the Open Language Archive Community (OLAC), a distributed network which does not have its own server but to which other repositories contribute. It is recognised as a good starting point for language resources in general and endangered languages in particular.

Dr Remijsen anticipated that interest in this dataset would increase with the passage of time and therefore the longer it is available, the better. The Dinka songs were recorded at the time when Sudan achieved independence and, therefore, a time of great social and political change.

3.2 Publication of research and data linking

As Dr Remijsen's research interests relate to language data, rather than song data, there are no planned publications related to them that would be authored by him. However, a range of research interests were focused on the project and other publications may appear.

3.3 Intended audience of the data

The data has a wide-ranging potential audience and capacity for further analysis in fields such as language, anthropology, sociology and ethno-musicology.

3.4 Reasons for the choices made in the 'data sharing matrix'

The general aim was to maximise the extent to which the material may be made publicly available on a not for profit basis. The informed consent gained from the participants in the course of the project was geared towards that outcome.

3.5 Conditions and constraints on the sharing of this project data

None; there was a desire on the part of the researchers not to maintain complex conditions of access which would ultimately go against the spirit of disseminating the data as widely as possible.

3.6 Requirement for data sharing usage statistics and measurements of use

Dr Remijsen acknowledged their usefulness, although he anticipated that it would probably be quite some time before statistics would emerge on future use of the data. He had no

strong views on the measurements of use beyond making the data as publicly available as possible.

3.7 Data sharing embargoes

As Dr Remijsen expressed it, minority languages are "a buyer's market", and he is thus happy to facilitate access to those researchers that are interested in making use of it.

3.8 Anticipated future use of the data

As stated above, the data has a wide-ranging potential audience and capacity for further analysis in fields such as language, anthropology, sociology and ethno-musicology.

3.9 Additional support services desired at the University of Edinburgh

The establishment of a standard procedure to archive and make available research datasets that result from the more regular studies that are undertaken within the School. At the moment such datasets simply remain on the researcher's computer or website once the resulting paper is published. Dr Remijsen also drew attention to the degree to which science journals in particular are now assuming some of this responsibility and offering researchers the opportunity to make their data available.

Data-sharing matrix: data types and levels of sharing anticipated

List each type of data here (planning documents, raw data, analysis, etc)	**Wouldn't share with anyone**	**Would share only with my collabor-ators**	**Would share with others in my field**	**Would share with other academics outside my field**	**Would share with the general public**
185 wav files: Dinka songs, and chunks of songs					X
127 TextGrid files: time-aligned annotations of the Dinka songs. This format is specific to the PRAAT software.					X
91 eaf files: the time-aligned annotations, converted from TextGrid to the ELAN format					X
137 metadata files (docx /doc / pdf): provide metadata per song		X			
1 wav file: Dinka song, not to be disseminated		X			
1 TextGrid file: time-aligned annotation of a Dinka song, not to be disseminated		X			

1 index file providing key information about each song (based on metadata, questionnaire and permissions documents); 1 readme file providing general background information					X
35 permission files (docx /doc / pdf) and 28 questionnaire files: both relate to specific singers		X			